芝　浦　工　業　大　学

博　士　学　位　論　文

数値計算の信頼性を保証する浮動小数点フィルタに関する研究

太田　悠暉

目次

表目次

図目次

記号・関数一覧

記号・関数	意味
$\mathbb{R}$	実数全体の集合
$\mathbb{Z}$	整数全体の集合
$\mathbb{N}$	零を除いた自然数全体の集合
$\mathbb{N}_0$	零を含めた自然数全体の集合
$\mathbb{F}$	IEEE 754 で規定された, ある固定された精度における, 浮動小数点数全体の集合 (正規化数・非正規化数・零の和集合)
$\mathbf{u}$	丸めの単位 (the roundoff unit)
$\mathbf{u}_N$	正規化数の正の最小数
$\mathbf{u}_S$	非正規化数の正の最小数
`realmax`	$\mathbb{F}$ に属する最大数
Inf	無限大 (浮動小数点システムにおいて, 絶対値が `realmax` より大きな値に対する表現)
NaN	非数 ($\sqrt{-1}$, $\log(-1)$ のような無効な演算に対する表現)
$\mathrm{fl}(\cdot)$	括弧内のすべての演算を, 最近点丸めを利用して, 浮動小数点演算を用いて評価した結果
$\mathrm{fl}_{\triangle}(\cdot)$	括弧内のすべての演算を, 上向き丸めを使用して, 浮動小数点演算を用いて評価した結果
$\mathrm{fl}_{\triangledown}(\cdot)$	括弧内のすべての演算を, 下向き丸めを使用して, 浮動小数点演算を用いて評価した結果
$\mathrm{float}(\cdot)$	括弧内のすべての演算を, 最近点丸めを利用し, 任意の順序で, 浮動小数点演算を用いて評価した結果 (演算子の優先順位が同じ演算のみ計算順序を変更してもよい)
$\mathrm{ufp}(a)$	$\lvert a\rvert$ 以下の最大の 2 のべき乗数を返す
$\mathrm{succ}(a)$	a より大きい最小の浮動小数点数を返す
$\mathrm{pred}(a)$	a より小さい最大の浮動小数点数を返す
$\mathrm{FMA}(a,b,c)$	$a\cdot b+c$ に最も近い浮動小数点数を返す $\mathrm{fl}(a\cdot b+c)$ であれば 2 演算となるが, $\mathrm{FMA}(a,b,c)$ では途中の結果を丸めず, 1 演算で評価される

1 序論

数値計算では, IEEE 754 規格 [1] に基づいた浮動小数点演算が用いられることが多い. しかし, 浮動小数点演算を用いた場合, 丸め誤差の影響により正確な値が得られないことがある. これは, 浮動小数点数は有限桁のビットを用いて数を表現する規則により, 2 項演算の結果が常に浮動小数点数となるとは限らないためである. このため, 数値計算結果に対する誤差の上限を求める方法は, 精度保証付き数値計算という分野で活発に研究されてきた [16, 18, 19, 25].

ここで, 数値計算によって誤判定する例をいくつか紹介する. ただし, 演算は IEEE 754 に基づく最近偶数丸めに従うものとし, $\mathbf{u}$ を相対丸めの大きさとする. 例えば, $1 + \mathbf{u}$ の計算結果は 1 となる. すなわち, 以下の式を前から順に数値計算で求めた結果も 1 となる.

$$((((1 + \mathbf{u}) + \mathbf{u}) + \mathbf{u}) + \cdots + \mathbf{u}) + \mathbf{u}.$$

これを拡張し, 長さ n のベクトル p を以下のように定める.

$$p_1 = 1, \quad p_i = \mathbf{u}, \quad p_{n-1} = -1, \quad p_n = -\mathbf{u}, \quad (2 \le i \le n-2).$$

$n \ge 4$ のとき, その総和 $\sum_{i=1}^{n} p_i$ を前から順に数値計算で求めると -2^{-53} となる. しかし, 真の結果は $(n-4)\mathbf{u}$ である. 特に $n \ge 5$ のときは, 真の結果と数値計算の符号が異なることになる.

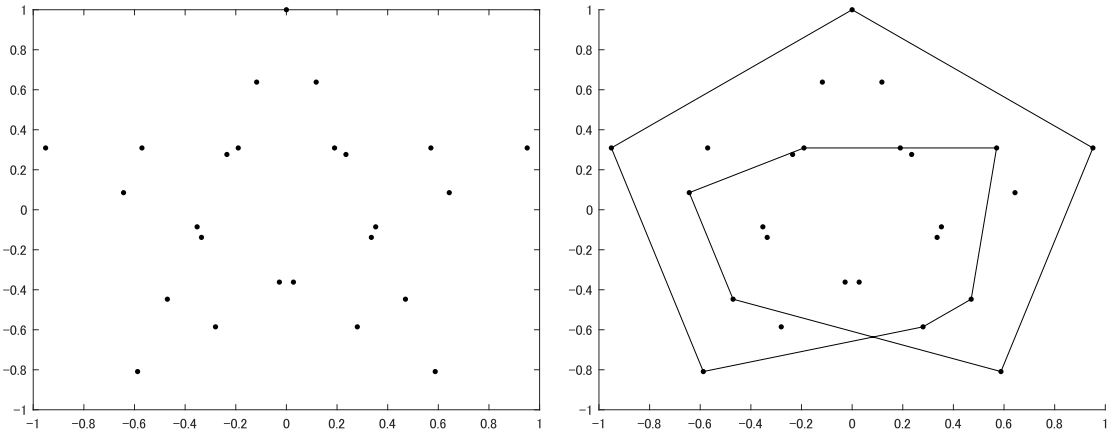
また, $1, -\mathbf{u}, 2\mathbf{u}+4\mathbf{u}^2 \in \mathbb{F}$ の和と $1, \mathbf{u}, \mathbf{u} \in \mathbb{F}$ の和の大小関係を考える. 前者は $1+\mathbf{u}+4\mathbf{u}^2$, 後者は $1 + 2\mathbf{u}$ であるため, binary32, binary64, binary128 では後者の方が大きい. しかし $(1 + (-\mathbf{u})) + (2\mathbf{u} + 4\mathbf{u}^2)$, $(1 + \mathbf{u}) + \mathbf{u}$ の順で数値計算により得られた結果はそれぞれ

$$1 + 2\mathbf{u}, \quad 1$$

となり, 正しい大小関係と異なる結果を得る. すなわち, このような単純な計算においても, 真値の大小関係と計算値の大小関係は異なることがある.

また, 丸め誤差の影響により, 凸包構成に失敗する例を紹介する. 2 次元における凸包は, 入力点のすべてを含む最小の凸多角形である. 図 1(a) のように配置した入力点に対して, 逐次添加法を利用して数値計算で求めた凸包は図 1(b) である. 得られた結果は明らかに凸包ではないことが分かる. これは, 凸包構成のアルゴリズムにおいて利用されている判定問題 (Orient2D) が, 数値計算の誤差によって判定間違いが何度も起こり, 不正確な結果を得た例である.

上述のように, 使用するアルゴリズムや計算式が正しくても, 意図した結果を得られない場合がある. 例として示した計算式のように, 真の結果が計算できる場合や凸包のように可視化することで結果が正しいかどうか判断することができる場合はよいが, そうでない場合は誤った判定を見逃してしまう可能性がある. その場合, 例えば分岐処理の判定を間違え



(a) 入力点　　(b) 数値計算結果により得られた正しくない凸包

図 1: 入力点とそれに対する近似計算を利用して得られた凸包

て, 本来実行されない処理が実行されてしまうことで, 最終結果に大きく影響を与えることがある. よって, 判定間違いを起こさないためには, 十分な精度の多倍長精度計算や, 厳密な計算を行える数式処理などが用いられる. しかし, それらの計算は数値計算に比べて計算速度が低速であることが知られている. また, 数値計算による判定結果はいつでも間違うわけではないため, 厳密な計算がいつでも必要であるとは限らない. そのため, すべての数値計算を誤差のない計算にそのまま置き換えるのではなく, 必要に応じて誤差のない計算を導入する考え方がある. 特に, 計算幾何学では浮動小数点フィルタ [6, 13, 17, 21, 22, 29] や加速法 [20, 30–32] などとして知られている. 浮動小数点フィルタと加速法は, 符号などの判定が正しいことの十分条件を高速に検証可能な形式として提案されている. 特に浮動小数点フィルタは, 数値計算の判定結果が正しいための十分条件を与える, 浮動小数点演算で検証可能な条件式である. これらにより, 判定が正しいかどうかを検証し, 保証されない場合には高精度計算や厳密計算が用いられる. 例えば, 大小判定の問題に対して浮動小数点フィルタを適用する場合は, 図 2 のようなフローチャートとなる.

ここで, 浮動小数点フィルタの例として Orient2D を紹介する. Orient2D は計算幾何学の基礎的な判定問題の 1 つで, よく現れる. まず, 2 次元平面上に与えられた 3 点 $A(a_x, a_y)$, $B(b_x, b_y)$, $C(c_x, c_y) \in \mathbb{R}^2$ を考える. ただし, 2 点 A, B は異なる点とする. 点 A と点 B をこの順に通過する向きのある直線 (有向直線 AB と呼ぶ) に対して, 点 C が左側にあるか, 右側にあるか, 直線 AB 上にあるかを判断する問題が Orient2D である (図 3 参照). 3 点の位置関係は, 次の行列式 $\det(G)$ の符号から明らかになる.

$$\det(G), \quad G := \begin{pmatrix} a_x & a_y & 1 \\ b_x & b_y & 1 \\ c_x & c_y & 1 \end{pmatrix}.$$

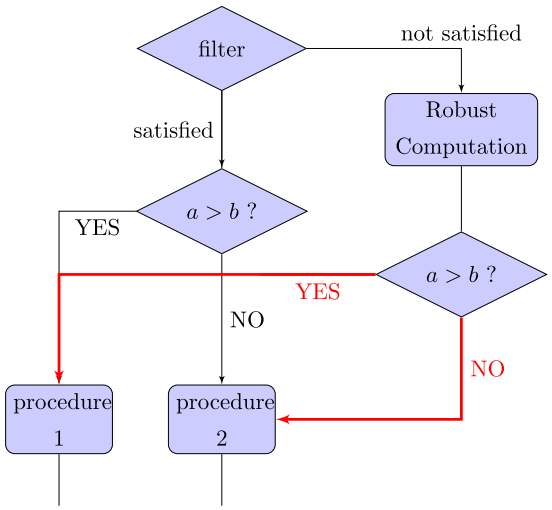


図 2: 浮動小数点フィルタによる判定のイメージ

具体的には行列式 $\det(G)$ の符号により, 次のように判断される.

$$\begin{cases} \det(G) > 0 & \Longleftrightarrow & \text{点 } C \text{ は有向直線 } AB \text{ の左側にある.} \\ \det(G) < 0 & \Longleftrightarrow & \text{点 } C \text{ は有向直線 } AB \text{ の右側にある.} \\ \det(G) = 0 & \Longleftrightarrow & \text{点 } C \text{ は直線 } AB \text{ 上にある.} \end{cases}$$

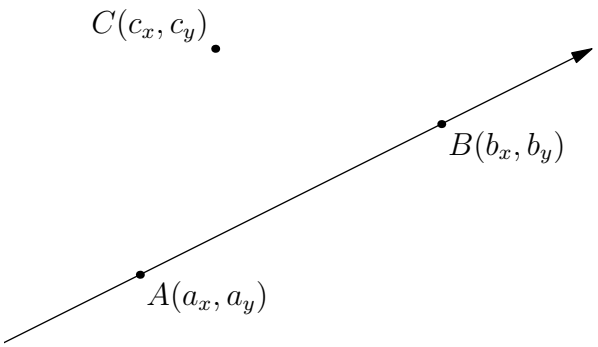


図 3: 点と直線の位置関係の判定問題 (Orient2D)

$\det(G)$ は三角形 ABC の符号付き面積とも呼ばれる.

3 点の位置関係を判断する際に, 数値計算では浮動小数点演算で評価した $\det(G)$ の近似値の符号を用いて判断する. しかし, 点 C が有向直線 AB に近接している場合や座標値の絶対値に大きな差がある場合には, $\det(G)$ の近似値の符号は, 丸め誤差の影響により $\det(G)$ の符号と異なる可能性がある. そこで, 丸め誤差を考慮した Orient2D の位置関係が, 数値計算のみで正確に判定可能であるかどうかを判断する不等式として, 浮動小数点フィルタが研究された [13, 29]. また, 一般の行列式に拡張して, その符号を判定する方法 [4, 23] や, 与えられた計算式から近似値の誤差上限の係数を簡単に求めるアルゴリズム [5] を利用した浮動小数点フィルタも提案されている. Orient2D に対する浮動小数点フィルタの一例として, 浮

動小数点例外を考慮したものを紹介する．これは浮動小数点フィルタがシンプルになるように, 丸め誤差解析が行われたものである．浮動小数点フィルタは, Algorithm 1 の 5 行目の判定部分となる．この浮動小数点フィルタ (不等式) を満たすとき, $\det(G)$ (Algorithm 1 では det) の符号と真の結果の符号が同じであることが保証される．ただし, $\theta, \mathbf{u}_N$ は定数である．

Algorithm 1 Orient2D の浮動小数点フィルタ [22]

Input: a_x, a_y, b_x, b_y, c_x, c_y: 各座標値

Output: det: 行列式

1: $l = (a_x - c_x) * (b_y - c_y)$
2: $r = (b_x - c_x) * (a_y - c_y)$
3: $det = l - r$
4: $errbound = \theta * (|l + r| + \mathbf{u}_N)$
5: **if** $|det| > errbound$ **then**
6: 　　return det
7: **end if**
8: % fall back to a more precise, slower method

Orient2D に限らず, 様々な誤差上限を評価したものが提案されているが, 目的の誤差評価式が提案されていない場合, または既存のものより改善した誤差評価を得たい場合, 個別に丸め誤差解析を行う必要がある．しかし, 丸め誤差を考慮しながら, その誤差上限を評価するのは大変な手間がかかる．そこで本論文では, 誤差評価が得られている場合に, 和・差および積に対する誤差上限を求める手法を提案する．これにより, 既存の誤差評価を組み合わせた誤差上限を求めることができる．また, 2 つの計算値の大小判定に対する浮動小数点フィルタも提案する．

本論文の構成は, 以下である．第 2 章において, IEEE 754 による浮動小数点数および浮動小数点演算を紹介する．また, それらの満たす性質, 本論文で使用する記号等についても紹介する．第 3 章において, 実数入力を考慮した Orient2D に対する浮動小数点フィルタを提案する．第 4 章において, 2 数の大小関係の判定に対する浮動小数点フィルタを提案する．第 5 章において, 符号のチェック, 絶対誤差・相対誤差に対する浮動小数点フィルタを提案する．また, 誤差評価が既知の 2 数の和および積を行った場合の誤差上限を求める方法も提案する．本論文において, 先行研究にない定理や補題について特に言及しない場合, 付録 A において, それらの証明を記述する．また, 参考文献と研究業績の表示および引用において, 区別しやすくするため参考文献は数字のみ, 研究業績は R を接頭文字とした数字で表記する．

2　記号・関数の定義

多くのコンピュータは有限桁の浮動小数点数を用いた浮動小数点演算を行う方式を採用している．これは, 実数を用いた実数演算（数学的に正しい演算）とは異なるため, 数値計算により得られた結果の精度を検証する必要がある．そこで本章において, まず IEEE の 1 部会が推奨する規格 IEEE 754 [1] に基づいた, 浮動小数点システムについて紹介する．その後に, 本論文で用いる記号や関数について紹介する．

2.1　IEEE 754 規格について

IEEE 754 とは, 浮動小数点数に関する表現と演算を規定した IEEE の規格である．IEEE 754 の 2019 年 2 月現在の正式な規格名は IEEE Standard for Floating-Point Arithmetic (ANSI/IEEE Std 754-2008) である．一般的には, IEEE std 754 や IEEE 754 と略記することが多い．また, 現在の IEEE 754 は 1985 年制定当初の規格 [2] を改訂したものであるため, IEEE 754-2008, IEEE 754-1985 と表記して区別することもある．本論文で特に断らずに IEEE 754 と表記した場合, IEEE 754-2008 の規格を指すこととする．

IEEE 754 は, 浮動小数点数を用いた計算で最も広く採用されている標準規格であり, 多くの CPU や FPU, ソフトウェアで実装されている．実際, 多くのプログラミング言語 (C, Fortran, Java など) において, 浮動小数点数処理の一部または全部として採用されている．ただし, Java では丸めについて IEEE 754-1985 を一部満たしていないため, 数値計算すると満足できる結果とはならないと Kahan は述べている [10].

2.2　浮動小数点数

一般に浮動小数点数は, 小数点の位置を変えない数の表現形式である固定小数点数に対して, 小数点の位置を移動させる数の表現形式である．浮動小数点数は非常に大きな値や小さな値などを表現する際に用いられ, コンピュータではよく使われている．実生活では, 数の大きさの単位 (1 万円や 42.195km, 1cm など) として用いている．例えば, アボガドロ数[*1] は, 固定小数点数で表現すると $602\,214\,085\,700\,000\,000\,000\,000\,\mathrm{mol}^{-1}$ である．これに対し, 浮動小数点数では $6.022\,140\,857 \times 10^{23}\,\mathrm{mol}^{-1}$ と表現するため, その大きさと有効桁数が分かりやすい．本節において IEEE 754 に基づく浮動小数点数の表現方法などを記述する．

[*1] アボガドロ数は N_A と表現する．2018 年 10 月時点の推奨値は $6.022\,140\,857(74) \times 10^{23}\,\mathrm{mol}^{-1}$ である．2017 special CODATA adjustment [15] では $6.022\,140\,758(62) \times 10^{23}\,\mathrm{mol}^{-1}$ とされており, 2018 年 11 月 16 日の Conférence générale des poids et mesures (CGPM) で N_A を含む定義改定が審議され, 新定義が採択された．新定義は 2019 日 5 月 20 日から適用されることも合わせて決定された．また, 括弧内の数値は最後の桁の標準不確かさ (標準偏差で表した測定結果の不確かさ) を表す．

2.2.1 IEEE 754 の浮動小数点システム

IEEE 754 に基づく浮動小数点システムは, 浮動小数点データ (floating-point datum) とそれらの演算によって定義されている. 浮動小数点データには, 正規化数 (Normalized Number)・非正規化数 (Subnormalized Number)・零 (Zero)・無限大 (Infinities)・非数 (Not a Number) の 5 種類がある. 正規化数・非正規化数・零・無限大はまとめて浮動小数点数 (floating-point number) と呼ぶ. 浮動小数点数のうち正規化数・零・非正規化数は, 符号部 S (Sign), 指数部 e (Exponent), 仮数部 T (Significand), 基数 b (base) として次のように定められている.

$$(-1)^{\mathrm{Sign}} \times \mathrm{base}^{\mathrm{Exponent}} \times \mathrm{Significand}$$

上式は S, b, e を用いて次のように表現される. ただし, 仮数部の桁数 $t = p - 1$ (精度 p) とする.

$$(-1)^S \times b^e \times \left(\frac{d_0}{b^0} + \frac{d_1}{b^1} + \frac{d_2}{b^2} + \frac{d_3}{b^3} + \cdots + \frac{d_{p-1}}{b^{p-1}} \right) \quad \text{ただし, } 0 \leq d_i < b$$

この形式で表現可能な正規化数は, 基数 b, 仮数部の桁数 t (精度 p), 指数部の最大値 $emax$ によって決定される. 基数 b が 2 であり, 次を満たす数全体を 2 進正規化数という.

- S は 0 または 1 である.
- $emin \leq e \leq emax$ を満たす. ただし, $emin = 1 - emax$ である.
- d_0 は 0 でない.

IEEE 754 では, 基数 b が 2 または 10 である場合が規定されている. それらは, binary32, binary64, binary128, decimal64, decimal128 の 5 種類の形式である. 基数 b が 2 の場合は表 2 のようにまとめられる. 数値計算では, 主に binary32, binary64 が用いられている.

表 2: IEEE 754 の標準浮動小数点数形式

形式名	一般名	基数 b	総桁数 $k=1+w+t$	指数部桁数 w	仮数部桁数 t, 精度 p	$emax$ $b^{w-1}-1$
binary32	2 進単精度	2	32	8	23, 24	127
binary64	2 進倍精度	2	64	11	52, 53	1023
binary128	2 進 4 倍精度	2	128	15	112, 113	16383

ただし, 指数部の e は正・零・負のいずれの値もとることができる. このままでは数値全体の符号部とは別に, 指数部においても符号を持つことになり, 単純な大小比較ができなくなる[*2]. そのため, 指数部はすべて正の値となるように特定の値を足して表わす. この特定の

[*2] 負の値は符号ビットを持たない場合, 補数を用いて表現するため, 補数を変換する必要がある. そのため, まず補数なのかどうかの判定を行い, 変換した後に大小比較を行うことになる.

値をバイアス (bias) といい, 指数部を $E = e + bias$ で表わす表現形式をバイアス表現という. $bias = emax$ とすることがすべての形式で求められている. 有限の数かどうかにかかわらず数を保存する場合には, 指数部はバイアス $(+bias)$ した値として保存する. 反対に, データを表示する際には指数部を逆バイアス $(-bias)$ して e を求めることになる. 以降では, バイアス表現したことによる指数部は “大文字” で, バイアス表現ではない指数部は “小文字” で表している. 正規化数の場合の E の範囲は, $1 \leq E \leq 2emax$ である.

2.2.2 浮動小数点データの表現

2.2.1 の内容を踏まえたうえで, 浮動小数点データを表現する方法が考えられている. 浮動小数点データを 2 進数表現で表した場合, 図 4 のようになる. ただし, MSB は most significand bit (最上位ビット), LSB は least significand bit (最下位ビット) の略である.

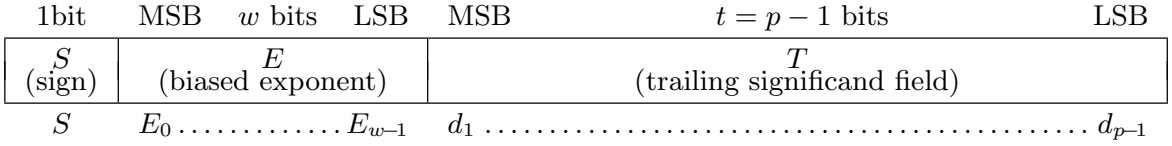


図 4: 基数 b が 2 である場合の浮動小数点数のビット表現 (IEEE 754-2008 Fig 3.1)

図 4 は, 符号部 S・指数部 E・仮数部 $T = d_1 d_2 d_3 \cdots d_{p-1}$ の順に繋げた次の形で表現されることを表している.

$$SE_0E_1E_2 \cdots\cdots\cdots E_{w-1}d_1d_2d_3 \cdots\cdots\cdots d_{p-1}$$

上式において d_0 を省略しているが, これは意味のある省略である. 正規化数・零・非正規化数の場合は, 指数部 E の値を利用して d_0 を決定する (2.3 節参照).

また, 基数 b が 2 でない場合も, 同じようにして定義されている.

2.3 2 進浮動小数点数

コンピュータ内部では 0 または 1 を用いて表現し, 演算を行う. そのため, 基数 b が 2 である, 2 進数表現を用いる場合を記述する. 2 進数表現の場合 $d_i = \{0, 1\}$ であり, このとき表すことができる浮動小数点データを 2 進浮動小数点数という. 2 進浮動小数点数のうち有限の値は, 次の形式で表せる値の全体である.

$$(-1)^S \times 2^e \times \left(\frac{d_0}{2^0} + \frac{d_1}{2^1} + \frac{d_2}{2^2} + \frac{d_3}{2^3} + \cdots + \frac{d_{p-1}}{2^{p-1}} \right).$$

2.3.1 正規化数

仮数部で d_0 を 1 と定めると, 数を一意に表現できる[*3]. このように, d_0 を 1 と定めることを浮動小数点数についての正規化といい, 正規化した数を正規化数という. また $b = 2$ の場合, 式 (2.3) において $d_0 = 1$ である数を 2 進正規化浮動小数点数 (2 進正規化数) という. 2 進正規化数は $d_i = \{0, 1\}$ に対し,

$$(-1)^S \times 2^e \times \left(\frac{1}{2^0} + \frac{d_1}{2^1} + \frac{d_2}{2^2} + \frac{d_3}{2^3} + \cdots + \frac{d_{p-1}}{2^{p-1}} \right) \quad \text{ただし, } emin \leq e \leq emax$$

と表せる数全体である.

ここで, 2 進正規化数で表わせる有限の数 x の絶対値の最大値 $x_{\max}$ は,

$$x_{\max} = 2^{emax} \times \left(\frac{1}{2^0} + \frac{1}{2^1} + \frac{1}{2^2} + \frac{1}{2^3} + \cdots + \frac{1}{2^{p-1}} \right)$$

である. binary64 では, $x_{\max} = 2^{1024} \times (1 - 2^{-53})$ である. また, 2 進正規化数で表わせる有限の数 x の絶対値の最小値 $x_{\min}$ は,

$$x_{\min} = 2^{emin} \times \left(\frac{1}{2^0} + \frac{0}{2^1} + \frac{0}{2^2} + \frac{0}{2^3} + \cdots + \frac{0}{2^{p-1}} \right) = 2^{emin}$$

である. binary64 では, $x_{\min} = 2^{-1022}$ である.

2.3.2 零

IEEE 754 では, 2 種類の零を規定しており, 次のように d_0 も含めて仮数部の d_i をすべて 0, 指数部を $E = 0$ $(e = emin - 1)$ として表わす.

$$(-1)^S \times 2^{emin-1} \times \left(\frac{0}{2^0} + \frac{0}{2^1} + \frac{0}{2^2} + \frac{0}{2^3} + \cdots + \frac{0}{2^{p-1}} \right).$$

符号部によって $+0$, -0 のどちらであるかを表現する. ただし, $+0 = -0$ と定めている.

2.3.3 非正規化数

浮動小数点システムの統一的な規格 (IEEE 754) が策定される前においては, $0 < |x| < x_{\min}$ を満たす数 x が与えられた場合, 多くのコンピュータは零と認識していた. そのため, 意図しない零による除算が起こりやすくなり, 無効な演算 (2.6 節参照) となることが多かった. そこで, 無効な演算が発生する割合を減らすため, $0 < |x| < x_{\min}$ となる数 x を表現する方法が考えられた. その方法により表現できる数を非正規化数[*4]という. 2 進非正規化浮

[*3] $d_0 = 1$ と決定しない場合, $1 \times 1.0 = 2 \times 0.5$ のように 2 通りで表現できるため一意ではない. 1×1.0 は $e = 0$, $d_0 = 1, d_i = 0$ $(i = 1, 2, 3, \cdots, p-1)$, 2×0.5 は $e = 1$, $d_1 = 1, d_i = 0$ $(i = 0, 2, 3, \cdots, p-1)$ とした場合である. そのため d_0 を決定することで, 一意に表現できる.

[*4] IEEE 754-1985 では Denormalized Number と紹介されている.

動小数点数 (2 進非正規化数) は, d_0 を 0, 指数部を $E = 0$ ($e = emin - 1$) として表わす.

$$(-1)^S \times 2^{emin-1} \times \left(\frac{0}{2^0} + \frac{d_1}{2^1} + \frac{d_2}{2^2} + \frac{d_3}{2^3} + \cdots + \frac{d_{p-1}}{2^{p-1}} \right) \quad \text{ただし, } \exists i \ d_i \neq 0$$

2.3.2 小節から, $d_1 = d_2 = \cdots = d_{p-1} = 0$ の場合は零を表すため, 仮数部のビット d_i がすべて 0 ではないときに非正規化数を表すことに注意する必要がある. ここで, 非正規化数で表現できる数は指数部が不変となるため, 仮数部の MSB である d_1 から順に 0 が連なることになる. そのため, 表現したい数の絶対値が小さければ小さいほど, 有効桁数の意味で精度が下がる. また, CPU を用いる計算において, 非正規化数を含んでいる場合, その計算速度が低下することも知られている.

2.4 丸め処理

2.3 節において浮動小数点数を紹介したが, 浮動小数点数は離散的な値である. そのため, 浮動小数点数を利用した浮動小数点演算 (近似計算) による結果と, (数学的に正しい) 実数演算による真の値との間に差が生じることがある. それらの誤差について, 端的に表現した文章として以下が挙げられる [11].

> "Significant discrepancies [between the computed and the true result] are very rare, too rare to worry about all the time, yet not rare enough to ignore."
>
> W.M. Kahan

これは丸め (rounding), 丸め処理 (round-off) による丸め誤差 (round-off error) が原因である場合がある. 計算幾何学の数値計算において, 浮動小数点演算による丸め誤差により, 実際とは大きく異なる結果を返してしまう例が報告されている [12, 24].

"丸め" とは, 実数 x' をそれに近い浮動小数点数 x で近似的に表現することをいい, "丸め処理" とは, 丸めにより浮動小数点数 x を決定する際の規則である. また, "丸め誤差" は丸めによって生じる誤差 $|x - x'|$ のことである. これらは, 複数の演算を含んだ数値計算後の値に対しても適用して使用される. 丸めの性質を利用した誤差解析は, 事前誤差解析・事後誤差解析に含まれ, 浮動小数点フィルタの作成などに使われる [19].

丸め処理は大きく最近点丸めと方向丸めの 2 種類に分けられる. 本節では, これらについて記述する.

2.4.1 最近点丸め

通常, 計算機は数値が与えられた際, その値に最も近い浮動小数点数を選び保存する. この丸め処理のことを最近点丸め (rounding-direction attributes to nearest) という. 最近点丸めは, 丸め処理する前の数値の状態によって保存する浮動小数点数を決定する. 最近点丸めは次の 2 つが定められている.

—— **偶数丸め：**

2 進数表現の際の最近点丸めのデフォルトとして設定することが要請されており, IEEE 754 では roundTiesToEven と表記している．これは最も近い浮動小数点数へ丸める形式である．ただし, 丸めたい値に最も近い浮動小数点数が 2 つ存在する場合, 仮数部の LSB が 0 である方の浮動小数点数を採用する．

—— **発散丸め：**

10 進数表現の際の最近点丸めのデフォルトとして設定することが要請されており, IEEE 754 では roundTiesToAway と表記している．偶数丸めと同様, 最も近い浮動小数点数へ丸める形式である．ただし, 丸めたい値に最も近い浮動小数点数が 2 つ存在する場合, 零よりも遠い方の浮動小数点数を採用する．また, IEEE 754-2008 から新たに採用された形式である．

最近点丸めを行う計算において偶数丸めを利用した演算を fl と書く．ここで, 偶数丸めにおいて最近の浮動小数点数が 2 つ存在する場合の 2 例を紹介する．ただし, 2 進単精度 (binary32) の場合とする．

- 浮動小数点数で表せる $x = 1,\ y = 2^{-24}$ に対し, これらの和 $x + y = 1 + 2^{-24}$ は, 浮動小数点数では表すことができない．ただし, この値は隣り合う 2 つの浮動小数点数 $1,\ 1 + 2^{-23}$ のちょうど中点である (表 3 参照)．そのため, 浮動小数点演算した結果 $\mathrm{fl}(x + y)$ は仮数部の LSB が 0 であるほうの値, すなわち 1 となる．
- 浮動小数点数で表せる $x = 1,\ z = 2^{-23} + 2^{-24}$ に対し, これらの和 $x + z = 1 + 2^{-23} + 2^{-24}$ は, 浮動小数点数では表すことができない．ただし, この値は隣り合う 2 つの浮動小数点数 $1 + 2^{-23},\ 1 + 2^{-22}$ のちょうど中点である．そのため, 浮動小数点演算した結果 $\mathrm{fl}(x + z)$ は仮数部の LSB が 0 であるほうの値, すなわち $1 + 2^{-22}$ となる．

表 3: 2 進単精度における偶数丸め

入力値	符号部 1 ビット	指数部 8 ビット	仮数部 23 ビット	仮想仮数部 24 ビット目	入力に対する丸め
$1+2^{-22}$	0	01111111	00000000000000000000010	0	なし
$1+2^{-23}+2^{-24}$	0	01111111	00000000000000000000001	1	上向き
$1+2^{-23}$	0	01111111	00000000000000000000001	0	なし
$1+2^{-24}$	0	01111111	00000000000000000000000	1	下向き
1	0	01111111	00000000000000000000000	0	なし

2.4.2 方向丸め

丸め処理する前の値は隣り合う 2 つの浮動小数点数の間にある場合が多い．その際, どちらの浮動小数点数を保存するかを使用するユーザーが決定することができる丸め処理が定められている．最近点丸めとは異なるこの丸め処理のことを方向丸め (directed rounding attributes) という．方向丸めは必要があれば設定でき, 次の 3 つの形式が定められている．

—— **上向き丸め：**
方向丸めの形式であり, IEEE 754 では roundTowardPositive と表記している．これは, 丸めたい値以上の浮動小数点数のうち最も小さい浮動小数点数を採用する．上向き丸めを行う計算において $\mathrm{fl}_{\triangle}$ と書く．

—— **下向き丸め：**
方向丸めの形式であり, IEEE 754 では roundTowardNegative と表記している．これは, 丸めたい値以下の浮動小数点数のうち最も大きい浮動小数点数を採用する．下向き丸めを行う計算において $\mathrm{fl}_{\triangledown}$ と書く．

—— **零向き丸め (チョップ)：**
方向丸めの形式であり, IEEE 754 では roundTowardZero と表記している．これは, その絶対値が丸めたい値の絶対値以下の浮動小数点数のうち, 最も丸めたい値に近い浮動小数点数を採用する．零向き丸めを行う計算において $\mathrm{fl}_{\bowtie}$ と書く．

以降の丸め処理は, 特に言及しない限り, IEEE 754 のデフォルトである最近偶数丸め (roundTiesToEven) とする．

2.5 無限大・非数

正規化数・零・非正規化数以外に浮動小数点データの特別な値として無限大・非数が用意されている．ここでは, これら 2 つについて記述する．無限大は厳密な数学の表現 (定義) とは異なるが, 浮動小数点数を用いた演算が破綻しないように設けられている．

2.5.1 無限大

数 x を丸めた値の絶対値が $x_{\max}$ 以下であれば, 正規化数・零・非正規化数のいずれかで表現される．丸めた値の絶対値が $x_{\max}$ を超える値は, 無限大を用い, Inf で表す．無限大は, 指数部を $E = 2emax + 1$ とし, $d_i = 0$ $(i = 1, 2, \cdots, p-1)$ で表現する．さらに $S = 0$ のときに $+$Inf, $S = 1$ のときに $-$Inf を表す．

2.5.2 非数

関数の定義域に存在しない値を関数に与えた場合, その関数値は数ではない．また, 数値計算途中に Inf となり, Inf $-$ Inf に対する結果も浮動小数点数で評価することはできない．

このように, 数ではない値を非数といい, NaN で表す. 非数には sNaN (Signaling NaN) と qNaN (Quiet NaN) の 2 種類が定められており, すべての浮動小数点形式で実装することが要請されている.

sNaN は, 不正演算例外として用いる. 例えば, 初期化されていない変数や無限大, 定義域以外の値を関数等で用いた際の演算結果として表現する. qNaN は, 計算機環境等によって異なるが, 無効なデータおよび結果から受け継いだ診断情報 (retrospective diagnostic information) を表現する.

演算結果が NaN となる場合, その種類 (sNaN か qNaN か) を決定するために必要な情報を, ビットパターンとして保持する. ビットパターンとして使用しない, 仮数部の残りのビットで診断情報を表現する. NaN の指数部は $E = 2emax + 1$ とすることが定められている. さらに, qNaN では $d_1 = 1$, sNaN では $d_1 = 0$ とすることが定められている. ただし, 無限大 (Inf) と区別するため, sNaN ($d_1 = 0$) の場合, すべての d_i $(2 \leq i \leq p-1)$ が 0 であってはならない. 仮数部の後ろの $p-2$ ビットの中で診断情報を表現する.

演算等により値が sNaN から qNaN になる場合は, d_1 を 0 から 1 に変更する. また, 通常符号部は意味を持たないが, 符号を必要とする演算がある場合にのみ意味を持つ. しかし, NaN は qNaN として保存する仕様になっていることが多い.

ここで, 浮動小数点形式の表現をまとめると表 4 のようになる.

表 4: 基数 b が 2 である場合の浮動小数点データの表現形式

Normalized Number	Subnormalized Number	Signed Zero	Inf	NaN
$1 \leq E \leq 2^w - 2$	$E = 0$	$E = 0$	$E = 2^w - 1$	$E = 2^w - 1$
$T : any$	$T \neq 0$	$T = 0$	$T = 0$	$T \neq 0$
$S = 0, 1$	$S = 0, 1$	$S = 0, 1$	$S = 0, 1$	$S = 0, 1$
←——————————Floating-Point Number——————————→				
←——————————————————Floating-Point Datum——————————————————→				

以降では, 浮動小数点数である正規化数・零・非正規化数・無限大全体の集合を $\mathbb{F}$ で表現する.

2.6 例外処理

正規化数・零・非正規化数で表現できない値や演算結果などは, 前節で紹介した無限大・非数で表現される. これらの結果となる演算等の処理を例外処理 (exception handling) という. ここでは, IEEE 754 が定めるデフォルトの例外処理について記述する.

2.6.1 無効な演算

定義域に含まれない値を入力とする演算は, 通常の計算としては定義されないが演算の評価は行われる. このような演算を無効な演算 (Invalid operation) といい, 結果は通常 NaN

になる．表 5 に例を示す．

表 5: 無効な演算となる例

演算例	結果
$0/0,\ 0\times\mathrm{Inf},\ \mathrm{Inf}/\mathrm{Inf},\ (+\mathrm{Inf})+(-\mathrm{Inf}),\ \sqrt{-1}$	NaN

2.6.2 零による除算

有限値の入力に対する演算の結果が, 無限大と定義されている場合がある．このような演算を Division by Zero という．例えば, Division by Zero は以下の場合に発生する．

- 有限の零でない数に対する除数が零
- 対数関数 $\log_b$ の真数が零

前者の場合, 分母の符号部と分子の符号部の排他的論理和が無限大の符号部になる．後者の場合, (無効な演算としての NaN ではなく) 無限大の符号はマイナスになる．

2.6.3 オーバーフロー

計算結果の絶対値が, $x_{\max}$ より大きい場合にのみ, オーバーフロー例外が示されることが要請されている．オーバーフローが発生した場合, その結果の符号は丸め処理および計算途中の結果の符号によって決定される．また, 丸め処理の違いによって結果も異なり, 以下のようになる．

a) 最近点丸め
   Inf または −Inf になる．ただし, 無限大の符号は計算途中の結果の符号と同じになる．
b) 下向き丸め
   $x_{\max}$ または負の無限大になる．
c) 上向き丸め
   $-x_{\max}$ または正の無限大になる．

零向き丸めの場合, $x_{\max}$ または $-x_{\max}$ となるため, オーバーフローは発生しない．

2.6.4 アンダーフロー

IEEE 754 では, アンダーフローについて次のように定めている．ただし, b^{emin} は正規化数で表せる有限の値の絶対値の最小数 $x_{\min}(=\mathbf{u}_N)$ である．

> 2 進形式の場合, 演算結果が丸め処理前・後にかかわらず, 零でない値に対して, 開区間 $(-b^{emin}, b^{emin})$ に包含されるとき, アンダーフローという例外状態であると表現することを要請する．

これより, “演算結果が零である場合は特例としてアンダーフローにはならない” ことが分かる. また, “演算結果が正規化数であっても, 計算途中にアンダーフローが起こっていないとは言えない” ことも読み取れる.

計算結果が正規化数であるにも関わらず, 計算途中にアンダーフローが起きる例として次の場合がある. 2 進倍精度 (binary64) の場合, $x = 2 - 2^{52} \in \mathbb{F}$, $y = 2^{-1023} \in \mathbb{F}$ の積 xy は $2^{-1022} - 2^{-1075}$ であるため, 浮動小数点演算の結果 $\mathrm{fl}(xy) = 2^{-1022}$ は正規化数 (の正の最小数) である. しかし, $-x_{\min} < xy < x_{\min}$ を満たすため, アンダーフローが起こっていることが分かる.

2.7 浮動小数点数の演算

ここまでは浮動小数点データの表現について紹介してきたが, ここではそれらを用いた演算について記述する. IEEE 754 において浮動小数点数を用いた演算は, 四則演算 (加減乗除) および根号の 5 種類のみが定められている. これらの演算について記述する. また, 解析に利用する丸め誤差評価についても紹介する. 本論文において, 特に言及しない限り, オーバーフローは発生しないと仮定する.

まず, 以下の記号と関数を紹介する.

- $\mathbb{F}$: IEEE 754 で規定された, 固定された精度における, 正規化数・非正規化数・零の和集合
- $\mathbb{U}$: IEEE 754 で規定された, 固定された精度における, 非正規化数と零の和集合
- $\mathrm{fl}(\dots)$: 括弧内のすべての二項演算を浮動小数点演算で評価した結果
- $\mathrm{float}(\dots)$: 括弧内のすべての二項演算を, 任意の順序で, 浮動小数点演算で評価した結果
- $\mathbf{u}$: 丸めの単位 (the roundoff unit)
- $\mathbf{u}_N$: 正規化数の正の最小数
- $\mathbf{u}_S$: 非正規化数の正の最小数

binary64 であれば, $(\mathbf{u}, \mathbf{u}_N, \mathbf{u}_S) = (2^{-53}, 2^{-1022}, 2^{-1074})$ である. 特に, 2 進浮動小数点数の場合, 以下の関係がある.

$$2\mathbf{u} \cdot \mathbf{u}_N = \mathbf{u}_S. \tag{2.1}$$

ここで, 関数 fl について補足する. $x, y \in \mathbb{F}$, $\circ \in \{+, -, \cdot\}$ に対して $\mathrm{fl}(x \circ y)$ は IEEE 754 における roundTiesToEven に従って計算値を返す. 演算毎に fl を用いると表記が長くなるため, 本論文では $\mathrm{fl}(\cdots)$ と表記した場合は, 括弧内のすべての演算を浮動小数点演算で評価するものとする. 例えば, $a, b, c \in \mathbb{F}$ に対して $\mathrm{fl}((a + b) + c)$ は $\mathrm{fl}(\mathrm{fl}(a + b) + c)$ を意味する. $\mathrm{float}(\dots)$ についても同様である.

また, $a \in \mathbb{R}$, $b, c, d \in \mathbb{F}$ に対して, 誤差解析に用いる関数を以下のように定義する.

$$\begin{aligned}
\mathrm{ufp}(a) &:= \begin{cases} 0 & \text{if } a = 0 \\ 2^{\lfloor \log_2 |a| \rfloor} & \text{otherwise} \end{cases}, \\
\mathrm{succ}(a) &:= \min\{x \in \mathbb{F} \mid x > a\}, \\
\mathrm{pred}(a) &:= \max\{x \in \mathbb{F} \mid x < a\}, \\
\mathrm{FMA}(b, c, d) &:= \min\{x \in \mathbb{F} \mid |x - (bc + d)|\}.
\end{aligned}$$

$\mathrm{ufp}(a)$ は, $a \neq 0$ のとき $|a|$ を超えない最大の 2 のべき乗数を求める関数である. $\mathrm{succ}(a)$ は a より大きい最小の浮動小数点数を, $\mathrm{pred}(a)$ は a より小さい最大の浮動小数点数を返す. FMA は通常 2 演算のところを, 積の結果を丸めずに和を求め, これに最も近い浮動小数点数を返す. fused multiply-add と呼ばれ, 丸め誤差の影響を 1 回に抑えることができる.

以後, 浮動小数点演算中の乗算について, 2 のべき乗数および定数との積を除き, 「$\cdot$」を用いて表記する. ただし, $\mathbf{u}$ と ufp がこの順で並ぶときは見やすさのために, 実数演算においても「$\cdot$」を用いることにする. この他, 添字つきどうしの積のように見づらい場合にも適宜「$\cdot$」を用いることにする. また, 使用する浮動小数点数は IEEE 754 の binary32, binary64, binary128 のいずれかとする.

これより, 浮動小数点数と浮動小数点演算についての定理をまとめる. 実数を浮動小数点数に丸めるとき, 以下の定理が知られている.

定理 2.1 ([6, 9, 17])
$x' \in \mathbb{R}$ と, x' を最近点の浮動小数点数に丸めた値 $x \in \mathbb{F}$ には以下の関係がある.

$$x' = (1 + \delta_{\mathrm{rnd}})x + \eta_{\mathrm{rnd}}, \qquad |\delta_{\mathrm{rnd}}| \leq \mathbf{u}, |\eta_{\mathrm{rnd}}| \leq \frac{\mathbf{u}_S}{2}.$$

x が正規化数であるとき $\eta_{\mathrm{rnd}} = 0$, 非正規化数または零のとき $\delta_{\mathrm{rnd}} = 0$ としてよい.

次に, 和・差および積の丸め誤差に関する定理を紹介する.

定理 2.2 ([6, 9, 17])
$x, y \in \mathbb{F}$ に対して, 次の関係が成り立つ.

$$\begin{aligned}
x \circ y = (1 + \delta_1)\mathrm{fl}(x \circ y) + \eta_{\mathrm{mul}}, \quad |\delta_1| \leq \mathbf{u}, \ |\eta_{\mathrm{mul}}| \leq \frac{\mathbf{u}_S}{2}, \quad \delta_1 \eta_{\mathrm{mul}} = 0, \\
|x \circ y| \leq (1 + \mathbf{u})\mathrm{fl}\left(|x \circ y|\right) + \frac{\mathbf{u}_S}{2}, \quad \circ \in \{+, -, \cdot\}.
\end{aligned}$$

ただし, $\circ \in \{+, -\}$ では $\eta_{\mathrm{mul}} = 0$ である. すなわち

$$|x \pm y| \leq (1 + \mathbf{u})\mathrm{fl}(|x \pm y|) \tag{2.4}$$

が成立する.

次に, ある浮動小数点数に対して, 2 のべき乗数をかけた場合に成立する不等式を紹介する.

補題 2.3

$a, b \in \mathbb{F}$, $b = \mathrm{ufp}(b)$ に対して

$$ab \leq \mathrm{fl}(a \cdot b) + \frac{\mathbf{u}_S}{2} \tag{2.5}$$

が成立する.

補題 2.3 の証明

アンダーフローが発生しなければ, 2 のべき乗数をかけても丸め誤差は発生せず, アンダーフローが発生する場合は (2.2) の δ_1 を 0 とすることにより得る. □

定理 2.2 は, 丸め誤差解析によく使われてきた. これに対して, Rump, Ogita, Oishi らは ufp を用いた丸め誤差解析モデルを提案した [28].

補題 2.4 ([28])

$x, y \in \mathbb{F}$ の和・差に対し, $\delta_2 \in \mathbb{F}$ が存在して以下を満たす.

$$\mathrm{fl}(x \circ y) = x \circ y + \delta_2, \quad |\delta_2| \leq \mathbf{u} \cdot \mathrm{ufp}(x \circ y) \leq \mathbf{u} \cdot \mathrm{ufp}(\mathrm{fl}(x \circ y)), \quad \circ \in \{+, -\}.$$

積に関して, $\delta_3, \eta \in \mathbb{R}$ が存在して, 以下を満たす.

$$\mathrm{fl}(x \cdot y) = xy + \delta_3 + \eta, \quad |\delta_3| \leq \mathbf{u} \cdot \mathrm{ufp}(\mathrm{fl}(x \cdot y)), \quad |\eta| \leq \frac{\mathbf{u}_S}{2}, \quad \delta_3 \eta = 0.$$

定理 2.5 ([27])

$p \in \mathbb{F}^n$ とする. このとき, 総和 $\sum_{i=1}^n p_i$ について, 以下の誤差評価が成り立つ.

$$\left| \sum_{i=1}^n p_i - \mathrm{float}\left(\sum_{i=1}^n p_i \right) \right| \leq (n-1)\mathbf{u} \cdot \mathrm{ufp}\left(\mathrm{float}\left(\sum_{i=1}^n |p_i| \right) \right). \tag{2.6}$$

ただし, 左辺と右辺の float の計算順は同じとする.

定理 2.5 より, 以下の補題が導ける.

補題 2.6

$p \in \mathbb{F}^n$ とする. このとき, 以下の評価が成り立つ.

$$\sum_{i=1}^n p_i \leq \mathrm{float}\left(\sum_{i=1}^n |p_i| \right).$$

2 つの浮動小数点数の差に関する補題を以下に紹介する.

補題 2.7 [R4]

$0 \leq a, b \in \mathbb{F}$ に対して, $a > b$ ならば $a \geq b + \mathbf{u} \cdot \mathrm{ufp}(a)$ が成立する.

ここで, $a \in \mathbb{F}$ のとき, ufp と succ に対して以下が成り立つ.

$$
\begin{aligned}
|a| &< 2\mathrm{ufp}(a) && a \neq 0, \\
|a| &\leq 2\mathrm{ufp}(a) - \max\{2\mathbf{u} \cdot \mathrm{ufp}(a), \mathbf{u}_S\} && a \neq 0, && (2.7)\\
\mathrm{succ}(a) &= a + \max\{2\mathbf{u} \cdot \mathrm{ufp}(a), \mathbf{u}_S\} && a \geq 0, && (2.8)\\
\mathrm{succ}(a) &\geq a + 2\mathbf{u} \cdot \mathrm{ufp}(a) && a \geq 0. && (2.9)
\end{aligned}
$$

補題 2.8 [R4]
$a \in \mathbb{R}$, $b \in \mathbb{F}$, $\circ = \{+, -, \cdot\}$ に対して以下が成り立つ.

$$
\begin{aligned}
&\mathrm{fl}(a) > b \Rightarrow a > b, \\
&\mathrm{fl}(a) < b \Rightarrow a < b. && (2.10)
\end{aligned}
$$

これは, 丸めの規則 (roundTiesToEven) より明らかである.

補題 2.9 [R4]
$0 \leq a, b \in \mathbb{F}$ に対して以下が成り立つ.

$$ab \leq \mathrm{fl}(\mathrm{succ}(a) \cdot b) + \frac{\mathbf{u}_S}{2}. \tag{2.11}$$

補題 2.10 [R4]
$0 \leq c, d < \mathbf{u}^{-1}$, $c, d \in \mathbb{F} \cap \mathbb{N}_0$ に対して, 以下が成り立つ.

$$c\mathbf{u} + d\mathbf{u}^2 \leq \mathrm{fl}(c\mathbf{u} + (d + \mathrm{ufp}(c))\mathbf{u}^2). \tag{2.12}$$

補題 2.11 [R4]
$0 \leq c, d < \mathbf{u}^{-1}$, $c \neq 0$, $c, d \in \mathbb{F}$, $c, d \in \mathbb{Z}$, $0 \leq f \in \mathbb{F}$ に対して, 以下が成り立つ.

$$(c\mathbf{u} + d\mathbf{u}^2)f \leq \mathrm{fl}((c\mathbf{u} + (d + 3\mathrm{ufp}(c))\mathbf{u}^2)f) + \frac{\mathbf{u}_S}{2}. \tag{2.13}$$

補題 2.12
$0 \leq c, d, e \in \mathbb{F}$ は整数, $0 \leq f \in \mathbb{F}$, $2 \leq k \leq \left\lfloor \dfrac{(1-\sqrt{2})(2c+7) + 2\sqrt{2(\mathbf{u}^{-1} - d - 4)}}{2\sqrt{2}} \right\rfloor$,
$const = 4 + 4k + \dfrac{k(k-1)}{2}$ とする. ただし, $k, const \in \mathbb{N} \cap \mathbb{F}$ とする. また, $c < \mathbf{u}^{-1}$, $d < \mathbf{u}^{-1}$ を仮定する. これらに対して, 以下が成り立つ.

$$(1 + \mathbf{u})^k \left(c\mathbf{u} + d\mathbf{u}^2\right) < c\mathbf{u} + \frac{1}{(1+\mathbf{u})^3} \left(kc + d + const\right) \mathbf{u}^2. \tag{2.14}$$

これより, 以下の 2 つの不等式が直ちに導かれる.

$$(1+\mathbf{u})^k\left(c\mathbf{u}+d\mathbf{u}^2\right) < \mathrm{fl}(c\mathbf{u}) + \mathrm{fl}\left((kc+d+const)\,\mathbf{u}^2\right), \tag{2.15}$$

$$(1+\mathbf{u})^k\left(c\mathbf{u}+d\mathbf{u}^2\right) < (1+\mathbf{u})\mathrm{fl}\left(c\mathbf{u}+(kc+d+const)\,\mathbf{u}^2\right). \tag{2.16}$$

3　実数入力を考慮した Orient2D

序論において, Orient2D に対する浮動小数点フィルタの先行研究について簡単に述べたが, それらは入力が浮動小数点数に限るものであった. すなわち, 入力点は誤差なく表現できているという暗黙の前提があった. ここでは, Algorithm 1 の拡張となるよう, 入力を長方形領域 (入力の実数の点を包含する領域) として捉えた場合の浮動小数点フィルタを作成する. なお, 本章の内容は研究業績 [R1, R5–R7, R13–R16, R18, R19] と関連がある.

本章においては, 入力値は実数であり, それらに最も近い浮動小数点数で表現されていると仮定する. すなわち, それぞれの入力値を a'_x, a'_y, b'_x, b'_y, c'_x, $c'_y \in \mathbb{R}$, それらに最も近い浮動小数点数を a_x, a_y, b_x, b_y, c_x, $c_y \in \mathbb{F}$ とする. このとき, 次のように表現できる.

$$a_x = \mathrm{fl}(a'_x),\ a_y = \mathrm{fl}(a'_y),\ b_x = \mathrm{fl}(b'_x),\ b_y = \mathrm{fl}(b'_y),\ c_x = \mathrm{fl}(c'_x),\ c_y = \mathrm{fl}(c'_y)$$

また, これらは定理 2.1 を利用して, 次のように評価できる.

$$\begin{array}{llll} a'_x = a_x + r_{ax} + \eta_1, & r_{ax} = \delta_{ax} \cdot a_x, & a'_y = a_y + r_{ay} + \eta_2, & r_{ay} = \delta_{ay} \cdot a_y, \\ b'_x = b_x + r_{bx} + \eta_3, & r_{bx} = \delta_{bx} \cdot b_x, & b'_y = b_y + r_{by} + \eta_4, & r_{by} = \delta_{by} \cdot b_y, \\ c'_x = c_x + r_{cx} + \eta_5, & r_{cx} = \delta_{cx} \cdot c_x, & c'_y = c_y + r_{cy} + \eta_6, & r_{cy} = \delta_{cy} \cdot c_y. \end{array} \tag{3.1}$$

また, 誤差解析の簡略化のため, 次のような表記を導入する.

$$\begin{array}{llll} p_1 := a_x - c_x, & q_1 := \mathrm{fl}(a_x - c_x), & r_1 := |a_x| + |c_x|, & s_1 := \mathrm{fl}(|a_x| + |c_x|), \\ p_2 := b_y - c_y, & q_2 := \mathrm{fl}(b_y - c_y), & r_2 := |b_y| + |c_y|, & s_2 := \mathrm{fl}(|b_y| + |c_y|), \\ p_3 := a_y - c_y, & q_3 := \mathrm{fl}(a_y - c_y), & r_3 := |a_y| + |c_y|, & s_3 := \mathrm{fl}(|a_y| + |c_y|), \\ p_4 := b_x - c_x, & q_4 := \mathrm{fl}(b_x - c_x), & r_4 := |b_x| + |c_x|, & s_4 := \mathrm{fl}(|b_x| + |c_x|), \\ p_5 := p_1 \cdot p_2, & q_5 := \mathrm{fl}(q_1 \cdot q_2), & r_5 := r_1 \cdot r_2, & s_5 := \mathrm{fl}(s_1 \cdot s_2), \\ p_6 := p_3 \cdot p_4, & q_6 := \mathrm{fl}(q_3 \cdot q_4), & r_6 := r_3 \cdot r_4, & s_6 := \mathrm{fl}(s_3 \cdot s_4). \end{array}$$

以後, 上記を前提として解析を行う.

まず, 行列式 $\det(G)$ の評価を考える. ただし, 演算回数の点から

$$\det(G) = (a_x - c_x)(b_y - c_y) - (a_y - c_y)(b_x - c_x) = p_1 \cdot p_2 - p_3 \cdot p_4 = p_5 - p_6 \tag{3.2}$$

と式変形したものについて誤差解析を行う. 以下では, 浮動小数点演算においてアンダーフローが起き得る場合を考える. 定理 2.2 より $p_i = (1 + \delta_i)q_i$, $(1 \le i \le 4)$ と式変形される. また, $M_1 := (1 + \delta_1)(1 + \delta_2)$, $M_2 := (1 + \delta_3)(1 + \delta_4)$ とおけば,

$$p_5 = p_1 \cdot p_2 = M_1 \cdot q_1 \cdot q_2, \quad p_6 = p_3 \cdot p_4 = M_2 \cdot q_3 \cdot q_4$$

となる. さらに, 定理 2.2 から $q_1 \cdot q_2 = (1 + \delta_5)q_5 + \eta_1$, $q_3 \cdot q_4 = (1 + \delta_6)q_6 + \eta_2$ であるため, $M_3 := M_1(1 + \delta_5)$, $M_4 := M_2(1 + \delta_6)$ とすれば

$$p_5 = M_3 \cdot q_5 + M_1 \cdot \eta_1, \quad p_6 = M_4 \cdot q_6 + M_2 \cdot \eta_2$$

を得る. 以上より, 次の評価が得られる.

$$\begin{aligned} p_5 - p_6 &= M_3 \cdot q_5 - M_4 \cdot q_6 + M_1 \cdot \eta_1 - M_2 \cdot \eta_2 \\ &= q_5 - q_6 + (M_3 - 1)\, q_5 - (M_4 - 1)\, q_6 + M_1 \cdot \eta_1 - M_2 \cdot \eta_2. \end{aligned}$$

ここで, 定理 2.2 より $q_5 - q_6 = (1+\delta_7)\mathrm{fl}(q_5 - q_6)$ と評価され, 以下を得る.

$$\begin{aligned} p_5 - p_6 &= (1+\delta_7)\mathrm{fl}(q_5 - q_6) + (M_3 - 1)\, q_5 - (M_4 - 1)\, q_6 + M_1 \cdot \eta_1 - M_2 \cdot \eta_2 \quad (3.3) \\ &= (1+\delta_7)\left(\mathrm{fl}(q_5 - q_6) + \frac{(M_3 - 1)\, q_5 - (M_4 - 1)\, q_6 + M_1 \cdot \eta_1 - M_2 \cdot \eta_2}{1+\delta_7}\right). \end{aligned}$$

次に, $(|a_x| + |c_x|)(|b_y| + |c_y|) + (|a_y| + |c_y|)(|b_x| + |c_x|) = r_1 \cdot r_2 + r_3 \cdot r_4 = r_5 + r_6$ に対する, 評価を考える. 上と同様に解析すると, 定理 2.2 より次を得る.

$$r_5 + r_6 = M_5 \cdot s_5 + M_6 \cdot s_6 + M_7 \cdot \eta_9 + M_8 \cdot \eta_{10}.$$

ただし, M_5, M_6, M_7, M_8 は $|M_5| \leq (1+\mathbf{u})^3$, $|M_6| \leq (1+\mathbf{u})^3$, $|M_7| \leq (1+\mathbf{u})^2$, $|M_8| \leq (1+\mathbf{u})^2$ を満たす実数とする. これより $r_5 + r_6$ は

$$r_5 + r_6 \leq (1+\mathbf{u})^3(s_5 + s_6) + (|\eta_9| + |\eta_{10}|)(1+\mathbf{u})^2$$

と評価される. さらに, 定理 2.2 から α, $\beta \geq 0$, α, $\beta \in \mathbb{F}$ のとき $\alpha + \beta \leq (1+\mathbf{u})\mathrm{fl}(\alpha + \beta)$ が成り立つため, 以下を得る.

$$r_5 + r_6 \leq (1+\mathbf{u})^4 \mathrm{fl}(s_5 + s_6) + \mathbf{u}_S (1+\mathbf{u})^2. \quad (3.4)$$

アンダーフローが起きなければ, $\eta_9 = \eta_{10} = 0$ であるから, (3.4) の右辺第 2 項は 0 としてよい. 同様に定理 2.2 を繰り返し利用することにより, $(r_1 + r_2) + (r_3 + r_4)$ に対して次が成り立つ.

$$\begin{aligned} (r_1 + r_2) + (r_3 + r_4) &\leq ((1+\mathbf{u})s_1 + (1+\mathbf{u})s_2) + ((1+\mathbf{u})s_3 + (1+\mathbf{u})s_4) \\ &\leq (1+\mathbf{u})^2(\mathrm{fl}(s_1 + s_2) + \mathrm{fl}(s_3 + s_4)) \\ &\leq (1+\mathbf{u})^3 \mathrm{fl}((s_1 + s_2) + (s_3 + s_4)). \end{aligned}$$

3.1 浮動小数点フィルタの作成

これより, 入力が実数である場合を考慮した, Orient2D の判定結果を保証する浮動小数点フィルタを作成する. まず, $\det(G') = \mathrm{fl}(\det(G)) + \Delta_1$ と式変形し, $|\Delta_1|$ の上限を浮動小数点演算により求め, Δ_2 とする. ただし Δ_2 は, G' から G への丸め誤差, $\mathrm{fl}(\det(G))$ とその計算で発生する丸め誤差の和の上限であり, $\mathrm{fl}(\det(G))$ と Δ_2 を用いて $\det(G')$ を含む区間を考える. この区間が 0 を含まなければ, 丸め誤差があったとしても, $\mathrm{fl}(\det(G))$ の符号には影響しない. すなわち, $\det(G')$ と $\mathrm{fl}(\det(G))$ の符号は同じである.

よって, 本論文において扱う浮動小数点フィルタとしての不等式は

$$\mathrm{fl}\,(|\det(G)|) \geq \Delta_2, \quad |\Delta_1|<\Delta_2 \in \mathbb{F}$$

となる. この不等式により, 符号が保証される組み合わせと, 保証されない組み合わせに分けることができる. 浮動小数点フィルタが成立し, $\det(G')$ と $\mathrm{fl}(\det(G))$ が正である場合の例 $(\mathrm{fl}(\det(G)) - \Delta_2 > 0)$ を図 5 に示す. 以下に, Δ_2 を具体的に定めた浮動小数点フィルタについて簡単にまとめる.

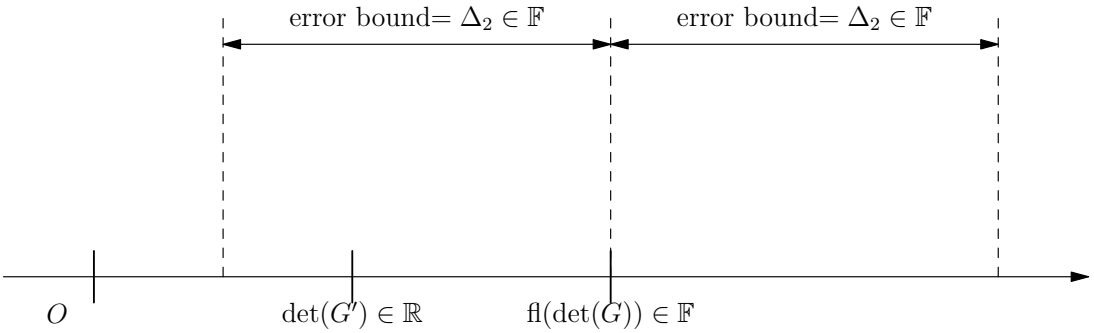


図 5: 浮動小数点フィルタのイメージ

定理 3.1 (実数入力に対する Orient2D の浮動小数点フィルタ) [R1]

入力値を丸めた値 a_x, a_y, b_x, b_y, c_x, c_y が正規化数・零・非正規化数のいずれかであるとする. このとき, 以下の不等式のうちいずれかを満たせば, $\det(G')$ と $\mathrm{fl}(\det(G))$ の符号は等しい.

$$\begin{aligned}&\mathrm{fl}\,(|q_5 - q_6|) &(3.6)\\ \geq&\mathrm{fl}((3\mathbf{u}+24\mathbf{u}^2)\cdot(|q_5| + |q_6|) + (2\mathbf{u}+24\mathbf{u}^2)\cdot(s_5 + s_6) + \mathbf{u}_N((s_1 + s_2) + (s_3 + s_4)) + \mathbf{u}_N),\\ &\mathrm{fl}\,(|q_5 - q_6|) \geq \mathrm{fl}((5\mathbf{u} + 40\mathbf{u}^2)\cdot(s_5 + s_6) + \mathbf{u}_N((s_1 + s_2) + (s_3 + s_4)) + \mathbf{u}_N). &(3.7)\end{aligned}$$

この浮動小数点フィルタは, アンダーフローが起きたとしても判定可能である.

定理 3.2 (正規化数に対する Orient2D の浮動小数点フィルタ) [R1]

入力値を丸めた値 a_x, a_y, b_x, b_y, c_x, c_y は正規化数のみであるとする. このとき, それぞれ以下の不等式のうちいずれかを満たせば, $\det(G')$ と $\mathrm{fl}(\det(G))$ の符号は等しい.

- 浮動小数点演算時にアンダーフローが起きない場合

$$\begin{aligned}&\mathrm{fl}\,(|q_5 - q_6|) \geq \mathrm{fl}((3\mathbf{u} + 16\mathbf{u}^2)\cdot(|q_5| + |q_6|) + (2\mathbf{u} + 20\mathbf{u}^2)\cdot(s_5 + s_6)), &(3.8)\\ &\mathrm{fl}\,(|q_5 - q_6|) \geq \mathrm{fl}((5\mathbf{u} + 32\mathbf{u}^2)\cdot(s_5 + s_6)). &(3.9)\end{aligned}$$

- 浮動小数点演算時にアンダーフローが起き得る場合

$$\begin{aligned}&\mathrm{fl}\,(|q_5 - q_6|) \geq \mathrm{fl}((3\mathbf{u} + 20\mathbf{u}^2)\cdot(|q_5| + |q_6|) + (2\mathbf{u} + 20\mathbf{u}^2)\cdot(s_5 + s_6) + \mathbf{u}_N), &(3.10)\\ &\mathrm{fl}\,(|q_5 - q_6|) \geq \mathrm{fl}((5\mathbf{u} + 32\mathbf{u}^2)\cdot(s_5 + s_6) + \mathbf{u}_N). &(3.11)\end{aligned}$$

表 6: 作成した浮動小数点フィルタの特徴

不等式	$a_x, a_y, b_x, b_y, c_x, c_y$ の条件	アンダーフローの発生
(3.6), (3.7)	なし	あり
(3.8), (3.9)	正規化数のみ	なし
(3.10), (3.11)	正規化数のみ	あり

特に binary64 では, 入力値を丸めた値すべての絶対値が 2^{-485} 以上ならば, 浮動小数点演算は計算途中も含め, 値は常に正規化数になる. そのため, この条件が満たされるときは (3.8), (3.9) を使用する. また, 表 6 にそれぞれの特徴をまとめた.

(3.6) の右辺は (3.7) の右辺よりも小さくなるが, (3.7) の右辺は (3.6) の右辺よりも計算のコストが小さい. (3.6) と (3.7) の右辺の差は $q_5 = q_6 = 0$ のとき最大で, 約 $3\mathbf{u}(s_5 + s_6)$ となる. そのため, (3.7) が成立しなくても, (3.6) では成立することもある. この他, (3.8) と (3.9), (3.10) と (3.11) の関係も同様である.

ここで, Bunikel, Funke, Seel らが提案したアルゴリズム[*5] [5] を利用した, Orient2D の浮動小数点フィルタは以下となる.

$$\mathrm{fl}(|q_5 - q_6|) \geq \mathrm{fl}((12\mathbf{u} + 16\mathbf{u}^2) \cdot (s_5 + s_6)). \tag{3.12}$$

(3.12) と (3.9) を比較すると, 提案するフィルタは係数について 2 倍以上過大評価を抑えていることが分かる.

Remark 1
Bunikel, Funke, Seel らが提案したアルゴリズムはさまざまな問題に対して, その誤差上限を複雑な誤差解析を行わずに求めることができる点が最大の特徴である. そのため, 同等の条件での比較ではないことに注意してほしい.

3.2 浮動小数点フィルタの凸包構成への応用

2 次元平面上における凸包を求めるアルゴリズムとして, Gift Wrapping, Graham's Scan, QuickHull, Incremental Convex Hull, Beneath-Beyond などがある. これらのアルゴリズムは [3, 20] などで紹介され, それらの内部では 3 点の位置関係の判定 (Orient2D) が必須の処理となっている. 前節で作成した浮動小数点フィルタを凸包構成のアルゴリズムに適用し, 反復的に凸包を再構成するアルゴリズムを提案する.

[*5] 対象となる式が四則演算および根号計算のみで構成されていた場合, 過大評価にはなるが, 計算結果の誤差上限を求める方法である.

3.2.1　凸包問題

平面上に与えられる n 個の点集合 S_n (図 6) に対し, S_n を含む最小の凸多角形 (図 7) を S_n の凸包 (convex hull) という [32]. さらに, 3 次元以上の高次元空間においても同様に凸包は定義される. この凸包は, 計算幾何学ではさまざまな場面 (ロボットが壁にぶつからずにゴールまでたどり着く導線の作成等) で用いられる.

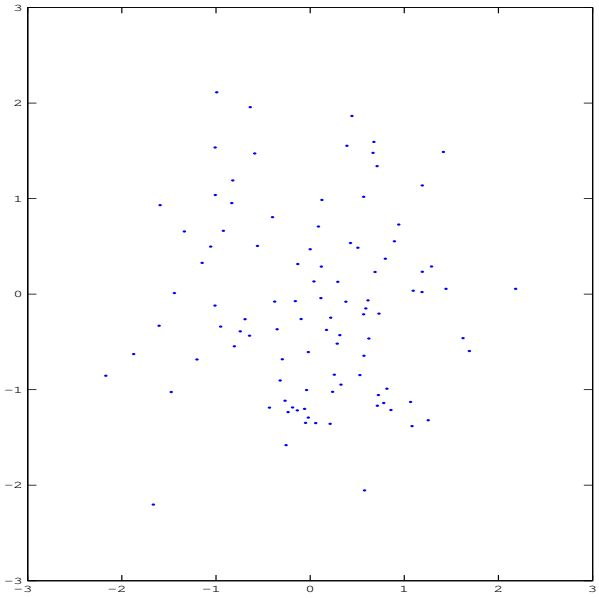


図 6: 点集合 S_n
$n = 100$ の場合

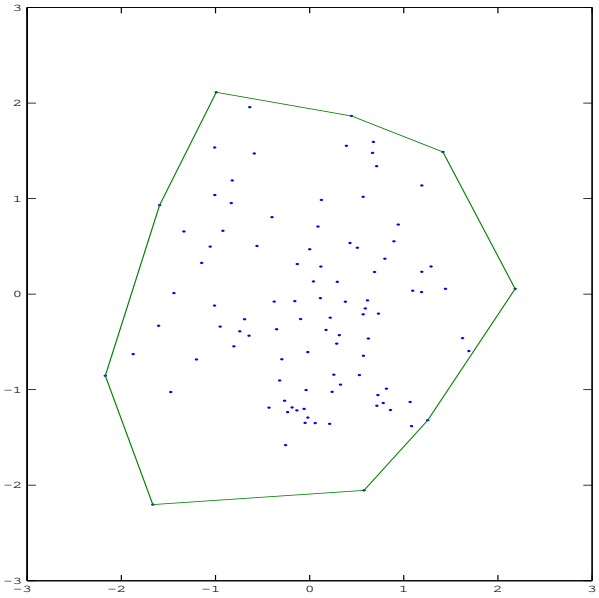


図 7: 点集合 S_n とその凸包の凸多角形
$n = 100$ の場合

3.2.2　逐次添加法と浮動小数点フィルタの適用の概略

凸包構成に対する逐次添加法は, 凸包に対して点を順次追加して, 凸包を再構成することで段階的に構成する, アルゴリズムの 1 つである. その手順は以下である.

- まず, n 個の点の中から凸包を構成する 3 点を選択し, それぞれ $p_0, p_1, p_2 (, p_0)$ とする. ただし p_0, p_1, p_2 の位置関係は, $\det(G)$ が正 (反時計回り) になるようにとる.
- 現在の凸包に, 点 p_i を追加して凸包を更新する.
- すべての点を追加して, 凸包を更新したら終了する.

上記で示した手順では, $\det(G)$ が正 (反時計回り) として紹介したが, $\det(G)$ が負 (時計回り) となるように選択してもよい. Algorithm 2 に, Orient2D の判定を利用した逐次添加法に基づく凸包アルゴリズムを記述する. 逐次添加法は, 凸包構成のアルゴリズムだけでなく, さまざまな問題に対して適用されている.

逐次添加法を利用して凸包を求めるアルゴリズム (Algorithm 2) に対して, 浮動小数点フィルタを応用する方法を以下に示す. はじめに, n 点の入力 p_i $(0 \leq i \leq n-1)$ を与える. 実数入力に対する凸包の頂点のインデックス列を, 浮動小数点演算により求めることを目的

Algorithm 2 逐次添加法に基づく凸包アルゴリズム (Orient2D による判定)

1. $\det(G) > 0$ となる 3 点を選択し, それらを p_0, p_1, p_2 とする.
2. CH は凸包の点列として, p_0, p_1, p_2, p_0 を初期値, $i = 3$ とする.
3. `While` $i \leq n-1$ `do`
   (a) 有向直線 $\mathrm{CH}_k\mathrm{CH}_{k+1}$ と p_i に対する $\det(G)$ を計算し, はじめて $\det(G) < 0$ となる CH_k に対する k を start, 再び $\det(G) > 0$ となる CH_k に対する k を finish とする.
   (b) 凸包の点列から start $< k <$ finish の CH_k を削除し, その間に p_i を追加して CH を更新する.

とする.

(i) 入力の n 点から $n-k$ 点に対する凸包を構成する. ただし, k はアルゴリズム中の浮動小数点フィルタによって, 有向直線との位置関係が保証されなかった点の数である.

(ii) 判定不能な k 点が, $n-k$ 点に対する凸包の内部または外部の点かどうかを Incremental Convex Hull Algorithm を用いて再確認する. もし外部の点と保証された場合は凸包を更新し, 凸包の内部または外部の点であると保証できなかった l 点 ($l \leq k$) を判定不能の点とする.

(iii) $n-l$ 点に対する凸包と, 判定不能な l 点を返す.

$k = l$ として (ii), (iii) を必要回数あるいは設定回数だけ繰り返す. もし $l = 0$ ($k = 0$) であれば, 所望の凸包が構成できていることになる. これにより "n 点に対する正しい凸包の点列", "部分的に正しい凸包の点列と判定不能の点列", "すべて判定不能の点列" のいずれかを得る.

次小節において, 凸包構成の精度保証アルゴリズムと (ii) において再確認する意義について述べる.

3.2.3 精度保証された逐次添加法による凸包構成のアルゴリズム

Incremental Convex Hull Algorithm は, k 点に対する凸包から $k+1$ 点に対する凸包を逐次的に構成するアルゴリズムである. 本論文では, p_0 を y 座標が最小の点とし, 反時計回りとなる凸包を構成する.

Incremental Convex Hull Algorithm に対して浮動小数点フィルタを適用した疑似アルゴリズムを, Algorithm 3 に示す[*6]. ただし, Algorithm 3 において, 下線は位置関係を決

[*6] 疑似アルゴリズムでは最端点を選択して内部の点を取り除くなどの前処理や, 分割統治法などを用いた手法については考慮しない. そのため, 最も単純な Incremental Algorithm である. ただし, 提案する浮動小数点フィルタは凸包構成のための前処理や, 他のさまざまな凸包構成のアルゴリズムへの応用も可能である.

定するための検証に必要であり, 追加した部分である. これらを取り除けば, Incremental Convex Hull Algorithm (Algorithm 2) と同じになる.

Algorithm 3 Verified Incremental Convex Hull Algorithm [R1]

Input : 凸包構成点列 CH, 点列 p

Output : 凸包構成点列 CH, 判定不能点列 v

Notation : CH_j は CH の j 番目の要素を表す

VIconvhull(CH, p)

$t = 0$, p の要素の個数を n とする

`for` $0 \leq i \leq n-1$

(a) CH の要素の個数を length, flag $= 0$ とする

(b) `for` $1 \leq j <$ length

i. $q_j = \mathrm{CH}_j$, $q_{j+1} = \mathrm{CH}_{j+1}$ とし, p_i を追加点とする

ii. q_j, q_{j+1}, p_i に対して行列式 fl(det(G)), 誤差の上限を計算する

iii. q_j, q_{j+1}, p_i に対して浮動小数点フィルタ (不等式) が成立する場合

flag == 0 かつ fl(det(G)) < 0 となったとき, start $= j$, flag $= 1$ とする

flag == 1 かつ fl(det(G)) > 0 となったとき, finish $= j$, flag $= 2$ とする

浮動小数点フィルタ (不等式) が成立しない場合

p_i を判定不能点として v_t に保存し, $t = t + 1$ とする

flag $= 0$ として j に対するループを抜ける

`end`

(c) (p_i が CH の外部にある場合 (flag $\neq 0$) に CH を更新する)

flag == 1 なら finish = length とする

start $< k <$ finish に対して CH_k を凸包構成点列から削除する. start と finish の間に p_i を追加する

`end`

`return` CH, v

3.2.2 節の (i)〜(iii) を実現するためのアルゴリズムを以下に示す (Algorithm 4).

提案したアルゴリズム (Algorithm 4) は図 8 のように, 判定不能点を複数回判定することによって, 凸包を構成する. ここで, 前節の (ii) において再確認する意義, すなわち Algorithm 4 の for 文について議論する. Incremental Algorithm においては, 個々の 3 点に対して浮動小数点フィルタが成立せずに判定不能であっても, 全体の凸包を構成する上では問題ない場合がある. 図 9 のような入力を例として与える. まず, p_0, p_1, p_2 を初期の凸包 CH(3) とする. 次に, p_3 を加えて CH(4) を構成する場合には, p_3 が直線 p_0p_1 に近接しているため判定不能となり, p_3 は除外され CH(4) は CH(3) と同じになる. しかし, CH(5) を構

Algorithm 4 Verified k-Iterative Incremental Convex Hull Algorithm [R1]

Input : 空の凸包構成点列 CH, 点列 p_i $(0 \leq i \leq n-1)$, 最大反復回数 N (≥ 1)

Output : 凸包構成点列 CH, 判定不能点列 p

VKconvhull(CH, p, N)

CH_1 を p のうち y 座標が最も小さい点をとり, 反時計回りになる相異なる 3 点を選択して CH(3) を構成し, p からそれらの点を取り除く

もし CH(3) を構成できなければ, すべて判定不能の点列として終了する

`for` $1 \leq k \leq N$

`if` length(p) $\neq 0$

（a）$lp = \mathrm{length}(p)$ とする

（b）[CH, p] = VIconvhull(CH, p)

% 関数 VIconvhull の結果を CH と p に代入

（c）$lp == \mathrm{length}(p)$ なら, k に対するループを抜ける

`end`

`end`

`return` CH, p

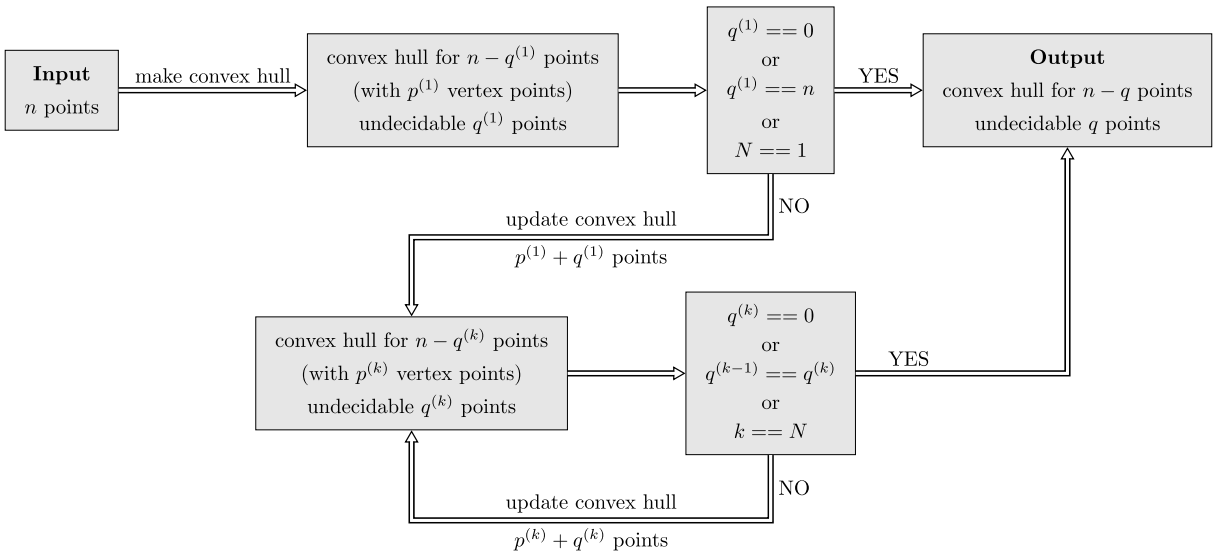


図 8: 提案するアルゴリズムのフロー

成した後に p_3 について再考すると, 容易に p_3 は CH(5) の内部であることが判定できる. このような状況を考慮するため, 1 度判定不能となった点に対して, 再確認することが有効である場合がある.

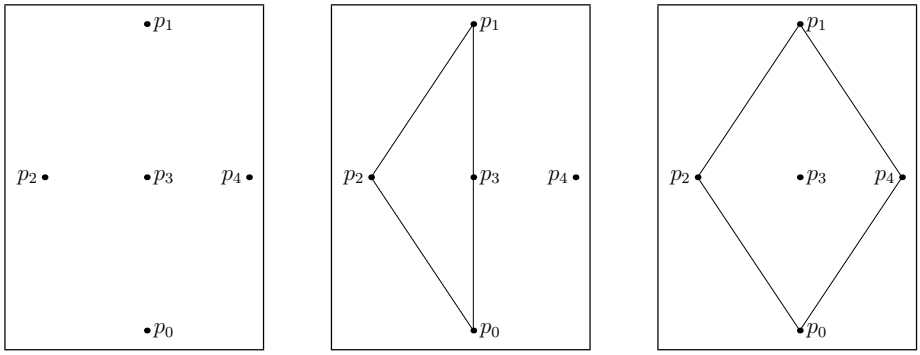


図 9: 提案するアルゴリズムにより凸包を構成するイメージ
左：入力 (5 点), 中央：CH(3) と CH(4), 右：CH(5)

入力に誤差がない場合には, 高精度計算を利用すれば 1 つ 1 つの辺に対して正しく判定できる. 図 9 であれば, p_0, p_1, p_3 の 3 点に対する位置関係も正しく判定される. しかし, 入力に誤差がある場合には座標値が正しくないため, 高精度計算による効果が見込めない. 提案するアルゴリズムは, このような問題に対応可能であり, さらに入力に誤差がない場合でも浮動小数点フィルタを応用することができる. ここで, 2 ページの図 1 で紹介した, 凸包構成に失敗する例に対して Algorithm 4 を適用した場合を示す. 1 回目で, 図 10 の左のように, 部分的に正しい凸包 (凸包の頂点である点 ∘ を実線で結んだ多角形) と浮動小数点フィルタを満たさなかった点 (∗ で示した判定不能点) を得る. 2 回目において, 判定不能点がないため, 正しい凸包 (図 10 の右) を得られたことが分かる. このように判定不能点がない, すなわち浮動小数点フィルタを満たさなかった点がない場合には, 数値計算のみで正しい凸包を得られたことが保証される.

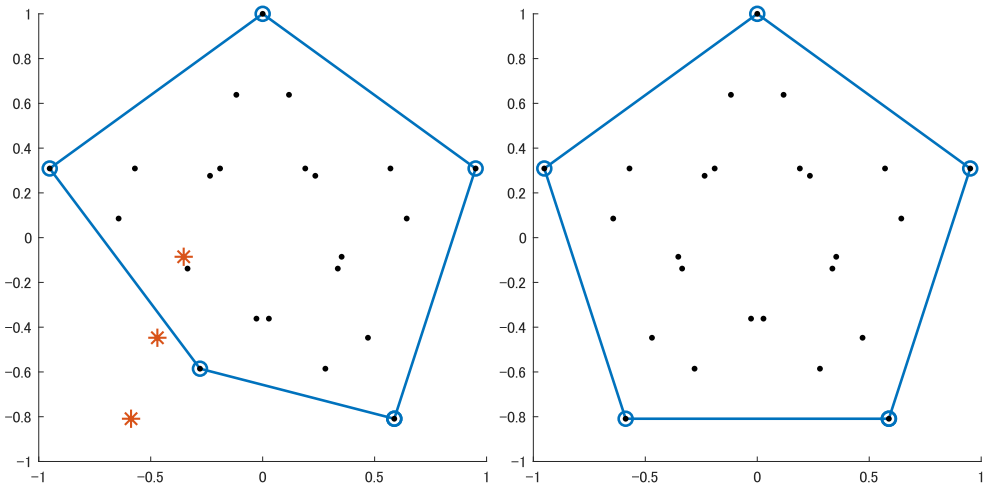


図 10: 判定不能点を再チェックすることで正しい凸包を得られる例

Remark 2

入力値である実数を保存できて実数演算を行える環境, あるいは入力値は多倍長浮動小数点数であるがより低い精度での計算が高速に行える環境であれば, 提案した浮動小数点フィルタ・アルゴリズムを使用することが可能である. まず, 浮動小数点演算により部分的に正しい凸包をできるだけ高速に構成し, 判定不能となった点について実数演算, あるいはもとの精度での浮動小数点演算, または高精度計算を用いてより正しい凸包を構成することができる. すなわち, 実数演算・高精度計算の前処理としても使用することができる.

4　2 つの計算値の大小判定に対する浮動小数点数フィルタ

本章では, 計算幾何学でよく用いられる浮動小数点フィルタの考え方を, 2 数の大小関係の判定に対して導入することを目的とする. ある計算における真の値を t_1, 浮動小数点演算による計算結果を x_1 とするとき, 丸め誤差に関する解析により

$$|x_1 - t_1| \leq (c_1\mathbf{u} + d_1\mathbf{u}^2)f + e_1\mathbf{u}_S, \quad c_1, d_1 \in \mathbb{N}_0, \quad 0 \leq f, e_1 \in \mathbb{F}, \tag{4.1}$$

$$|x_1 - t_1| \leq (c_2\mathbf{u} + d_2\mathbf{u}^2)f' + e_2\mathbf{u}_S, \quad c_2, d_2 \in \mathbb{N}_0, \quad 0 \leq f' \in \mathbb{R}, \quad 0 \leq e_2 \in \mathbb{F} \tag{4.2}$$

の形をした誤差上限を得ることが多くある[*7]. 例えば, ベクトルの総和と内積なら [6,26,27], ホーナー法による多項式の評価なら [8] などがある. 一般に浮動小数点フィルタを作成するには, 各演算に対する丸め誤差解析を利用し, 誤差上限を数値計算可能な形で表現する必要がある. その際, 評価式ごとに煩雑な丸め誤差解析が必要となる. ここでは, 先行研究の丸め誤差解析の結果を利用しつつ, 2 つの浮動小数点演算による計算結果から真の値の大小関係を保証する方法を提案する. なお, 本章の内容は研究業績 [R2, R8–R11, R17] と関連がある.

4.1　2 つの計算値の大小関係を保証する浮動小数点フィルタ

本節では, 2 数の大小判定に対する浮動小数点フィルタを作成する. ただし, (4.1) と同じ誤差評価が得られていると仮定する. すなわち, $t_i, s_i \in \mathbb{R}$, $x_i, e_i \in \mathbb{F}$, $0 \leq f_i \in \mathbb{F}$, $c_i, d_i \in \mathbb{N}_0$ $(i = 1, 2)$ に対して

$$t_i = x_i + \delta_i + s_i, \quad |\delta_i| \leq (c_i\mathbf{u} + d_i\mathbf{u}^2)f_i, \quad 0 \leq |s_i| \leq e_i\mathbf{u}_S, \quad c_i, d_i < \mathbf{u}^{-1} \tag{4.3}$$

と表現できるとする.

以下の補題により, (4.2) の形の誤差評価を得ていても, 適切に変換を行えば, 以降の議論を適用することができる.

補題 4.1 [R2]
(4.2) の誤差評価は, $0 \leq c_2, d_2 < \mathbf{u}^{-1}$ であるとき, (4.1) の形式に評価し直すことができる.

丸め誤差解析の結果を利用して $t_1 - t_2$ の符号と $\mathrm{fl}(x_1 - x_2)$ の符号が同じであることを保証する浮動小数点フィルタを作成する. t_1, t_2, x_1, x_2 に対して (4.3) を仮定する.

$x_1 - x_2$ と $\mathrm{fl}(x_1 - x_2)$ の差を δ とする. すなわち

$$x_1 - x_2 = \mathrm{fl}(x_1 - x_2) + \delta \tag{4.4}$$

[*7] (4.2) における f' の係数として $\gamma_n = \frac{n\mathbf{u}}{1-n\mathbf{u}}$, $(n < \mathbf{u}^{-1}, n \in \mathbb{N})$ を用いた丸め誤差解析も多く存在する [6,14,26]. ただし, $\mathbf{u}$ の項に対して $\mathbf{u}^3$ の項は相対的に小さいため, $\mathbf{u}^2$ の項までで上限を取っても, 多くの場合, 実用上大きな影響はない.

とする. 仮定した (4.3) および (4.4) より $t_1 - t_2$ は次のように変形できる.

$$t_1 - t_2 = (x_1 + \delta_1 + s_1) - (x_2 + \delta_2 + s_2) = \mathrm{fl}(x_1 - x_2) + \delta_1 - \delta_2 + s_1 - s_2 + \delta.$$

これより,

$$|x_1 - x_2| > |\delta_1 - \delta_2 + s_1 - s_2| \tag{4.5}$$

が成立すれば, $t_1 - t_2$ と $x_1 - x_2$ の符号は同じであり,

$$\mathrm{fl}(|x_1 - x_2|) > |\delta_1 - \delta_2 + s_1 - s_2 + \delta| \tag{4.6}$$

が満たされれば, $t_1 - t_2$ と $\mathrm{fl}(x_1 - x_2)$ の符号は同じである. 浮動小数点フィルタを作成するため, 2 つの補題を示す.

補題 4.2 [R2]
$c_i, d_i, e_i, f_i, \delta_i, s_i$ $(i = 1, 2)$ は (4.3) を満たすとする. また φ, α を

$$\varphi \in \{f \in \mathbb{F} \mid (1 + \mathbf{u})^3 \alpha \leq f\}, \quad \alpha := \max_{i=1,2} \left(c_i \mathbf{u} + d_i \mathbf{u}^2\right) \tag{4.7}$$

と定める. このとき,

$$|\delta_1 - \delta_2 + s_1 - s_2| < \mathrm{fl}\left(\varphi \cdot (f_1 + f_2) + (1 + 6\mathbf{u}) \cdot ((e_1 + e_2) + 1)\mathbf{u}_S\right) \tag{4.8}$$

が成立する.

φ についての 1 つの候補は $\mathrm{fl}\left(\max_{i=1,2}(c_i \mathbf{u} + (4c_i + d_i + 27)\mathbf{u}^2)\right)$ となる.

補題 4.3 [R2]
$c, d \in \mathbb{N}$, $0 \leq c, d < \mathbf{u}^{-1}$ に対して φ を

$$\varphi = \mathrm{fl}\left(c\mathbf{u} + (4c + d + 27)\,\mathbf{u}^2\right)$$

とすると, $(1 + \mathbf{u})^3 (c\mathbf{u} + d\mathbf{u}^2) < \varphi$ を満たす.

定理 4.4 [R2]
$t_i, x_i, c_i, d_i, e_i, f_i$ $(i = 1, 2)$ は (4.3) を満たすとする. また φ を (4.7) のように定める. このとき,

$$\mathrm{fl}(|x_1 - x_2|) > \mathrm{fl}\left(\varphi \cdot (f_1 + f_2) + (1 + 6\mathbf{u}) \cdot ((e_1 + e_2) + 1)\mathbf{u}_S\right) \tag{4.9}$$

が満たされるのならば, $\mathrm{fl}(x_1 - x_2)$ と $t_1 - t_2$ の符号は同じである.

定理 4.4 は x_1, x_2 を浮動小数点数と仮定し, アンダーフローのみを考慮していた. ここから x_1, x_2 自体が $\pm\mathrm{Inf}, \mathrm{NaN}$ であった場合や, 浮動小数点フィルタに必要な計算の中でオーバーフローや無効な演算が発生した場合について考察する.

定理 4.5 [R2]
x_i が $\pm\mathrm{Inf}, \mathrm{NaN}$ のいずれかならば, f_i は $\mathrm{Inf}, \mathrm{NaN}$ のどちらかになることを仮定する $(i = 1, 2)$. $x_1, x_2 \in \mathbb{F}$ のとき, (4.3) を仮定する. p_1 と p_2 を

$$p_1 := \mathrm{fl}(|x_1 - x_2|) \in \mathbb{F} \cup \{\pm\mathrm{Inf}, \mathrm{NaN}\},$$
$$p_2 := \mathrm{fl}\left(\varphi \cdot (f_1 + f_2) + (1 + 6\mathbf{u}) \cdot ((e_1 + e_2) + 1)\mathbf{u}_S\right) \in \mathbb{F} \cup \{\mathrm{Inf}, \mathrm{NaN}\}$$

とする. このとき, 論理式 $p_1 > p_2$ が真であれば, $\mathrm{fl}(x_1 - x_2)$ の符号は $x_1 - x_2$ の符号と等しい.

定理 4.5 の対偶より, 浮動小数点例外（オーバーフロー・無効な演算）が発生し得ると仮定しても, 大小関係が間違っているときには, 符号が同じであると保証することはない. また, ベクトルの総和では誤差上限を求める際, 評価式においてそれぞれ絶対値を用いるため, 定理 4.5 の仮定 "x_i が $\pm\mathrm{Inf}, \mathrm{NaN}$ のいずれかならば, f_i は $\mathrm{Inf}, \mathrm{NaN}$ のどちらかになる" を満たす. この他, 和・差・積で構成された計算式の場合, この仮定を満たす.

Remark 3
2 つの計算式があり, (4.3) を満たす誤差評価を得ていれば, その大小関係を判定する浮動小数点フィルタを作成することも可能である. 例えば, 内積の計算値と多項式の計算値の大小関係を保証する浮動小数点フィルタも簡単に作成できる.

Remark 4
(4.9) に関して, 浮動小数点数演算の内容 $(c_1, c_2, d_1, d_2, e_1, e_2)$ が事前にわかっている場合には, φ や $\mathrm{fl}\left((1 + 6\mathbf{u}) \cdot ((e_1 + e_2) + 1)\mathbf{u}_S\right)$ はプログラム中で計算せず定数として与えられる. この場合, 浮動小数点フィルタに必要な計算コストは

$$\mathrm{fl}(|x_1 - x_2|), \quad \mathrm{fl}\left(\varphi \cdot (f_1 + f_2) + (1 + 6\mathbf{u}) \cdot ((e_1 + e_2) + 1)\mathbf{u}_S\right)$$

にある足し算（引き算）3 回, 掛け算 1 回, 絶対値を 1 回取り, 比較を 1 回するのみである.

Remark 5
定理 4.4 では非正規化数 $\mathbf{u}_S$ を含む項がある. 演算に使用される数に非正規化数が含まれており, CPU で数値計算を行う際, 計算は非常に低速である場合がある. よって $\mathrm{fl}(e_1 + e_2) \leq \frac{1}{2}\mathbf{u}^{-1} - \frac{7}{2}$ のとき,

$$\mathrm{fl}\left((1 + 6\mathbf{u}) \cdot ((e_1 + e_2) + 1)\mathbf{u}_S\right) \leq \mathbf{u}_N$$

であるため, $\mathrm{fl}\left((1 + 6\mathbf{u}) \cdot ((e_1 + e_2) + 1)\mathbf{u}_S\right)$ を正規化数 $\mathbf{u}_N$ で置き換えることにより高速化を図ってもよい.

補題 4.6 [R2]
$\mathrm{fl}(e_1 + e_2) \leq \frac{1}{2}\mathbf{u}^{-1} - \frac{7}{2} \in \mathbb{F}$ のとき, $\mathrm{fl}\left((1 + 6\mathbf{u}) \cdot ((e_1 + e_2) + 1)\mathbf{u}_S\right) \leq \mathbf{u}_N$ となる.

4.2 浮動小数点フィルタの応用例

総和, 内積, ホーナー法で計算した 2 つの値の比較に対する浮動小数点フィルタを紹介する. float$(\cdots)$ を, 括弧内に存在する演算は浮動小数点演算とし, 任意の計算順による計算結果を意味する [8]. 例えば, $p \in \mathbb{F}^3$ に対して

$$\left| \mathrm{float}\left(\sum_{i=1}^{3} p_i\right) - \sum_{i=1}^{3} p_i \right| \leq \alpha$$

と記述すれば,

$$\left| \mathrm{fl}(\mathrm{fl}(p_1+p_2)+p_3) - \sum_{i=1}^{3} p_i \right| \leq \alpha, \quad \left| \mathrm{fl}(\mathrm{fl}(p_1+p_3)+p_2) - \sum_{i=1}^{3} p_i \right| \leq \alpha,$$
$$\left| \mathrm{fl}(\mathrm{fl}(p_2+p_3)+p_1) - \sum_{i=1}^{3} p_i \right| \leq \alpha$$

のすべてが成立することを意味する. ただし, 内積について float$(\cdots)$ を使用した場合, Winograd の方法 [33] など特別な計算順は採用されないとする.

4.2.1 総和に対する応用例

2 つの総和の比較についての浮動小数点フィルタの例を示す. $p \in \mathbb{F}^n$ とする. このとき, 総和 $\sum_{i=1}^{n} p_i$ について, 以下の誤差評価が知られている [27].

$$\left| \sum_{i=1}^{n} p_i - \mathrm{float}\left(\sum_{i=1}^{n} p_i\right) \right| \leq (n-1)\mathbf{u} \cdot \mathrm{ufp}\left(\mathrm{float}\left(\sum_{i=1}^{n} |p_i|\right)\right). \tag{4.10}$$

ただし, 左辺と右辺の float$(\cdots)$ の計算順は同じとする. ここで float$(\sum_{i=1}^{n} p_i)$ の結果が $\pm\mathrm{Inf}, \mathrm{NaN}$ のいずれかであれば, float$(\sum_{i=1}^{n} |p_i|)$ の結果は Inf であるため, 定理 4.5 の仮定は成立している[*8]. $x \in \mathbb{F}^m, y \in \mathbb{F}^n$ とし, float$(\sum_{i=1}^{m} x_i)$ と float$(\sum_{i=1}^{n} y_i)$ の大小関係から真の関係を保証する浮動小数点フィルタは,

$$c_1 = m-1, \quad d_1 = 0, \quad e_1 = 0, \quad f_1 = \mathrm{ufp}\left(\mathrm{float}\left(\sum_{i=1}^{m} |x_i|\right)\right),$$
$$c_2 = n-1, \quad d_2 = 0, \quad e_2 = 0, \quad f_2 = \mathrm{ufp}\left(\mathrm{float}\left(\sum_{i=1}^{n} |y_i|\right)\right)$$

として定理 4.4 を使用すればよい. 補題 4.3 を利用すれば, $\ell = \max(c_1, c_2)$, $\varphi = \mathrm{fl}(\ell\mathbf{u} + (4\ell+27)\mathbf{u}^2)$ である. よって,

$$\mathrm{fl}\left(\left|\mathrm{float}\left(\sum_{i=1}^{m} x_i\right) - \mathrm{float}\left(\sum_{i=1}^{n} y_i\right)\right|\right) > \mathrm{fl}\left(\varphi \cdot (f_1+f_2) + (1+6\mathbf{u})\mathbf{u}_S\right) \tag{4.11}$$

[*8] 4.2.2 節, 4.2.3 節で扱う丸め誤差評価に関しても, 同様の仮定は成立するため, 以後は特に断りを入れない.

が成立するならば, $\sum_{i=1}^{m} x_i$ と $\sum_{i=1}^{n} y_i$ の大小関係は近似計算の結果から保証される.

(4.10) の評価を利用することで得られた浮動小数点フィルタ (4.11) の φ が過大評価であることが懸念される. しかし, (4.10) の float の評価を fl にした場合, 浮動小数点フィルタ (4.11) の φ を $\varphi - 2\mathbf{u}$ にすることはできない. これについては, 後述する.

4.2.2 内積に対する応用例

2 つの内積値の比較を行う際の浮動小数点フィルタの例を示す. ここで, ベクトルに対する絶対値は, 各成分の要素について絶対値が作用するものとする. すなわち, $x \in \mathbb{R}^n$ に対して $|x| = (|x_1|, |x_2|, |x_3|, \ldots, |x_n|)^T$ である. $x, y \in \mathbb{F}^n$ に対する内積 $x^T y$ について, (4.1) に対応する以下の誤差評価が知られている [7].

$$|x^T y - \mathrm{float}(x^T y)| \le (n+2)\mathbf{u} \cdot \mathrm{ufp}\left(\mathrm{float}(|x|^T |y|)\right) + \frac{n}{2}\mathbf{u}_S, \quad (n+2)\mathbf{u} \le 1.$$

ここで, 左辺と右辺の $\mathrm{float}(\cdots)$ の計算順は同じとする.

$p, q \in \mathbb{F}^m$, $z, w \in \mathbb{F}^n$ とする. $p^T q$ と $z^T w$ に関して

$$\begin{aligned} c_1 = m+2, \quad d_1 = 0, \quad e_1 = \frac{m}{2}, \quad f_1 = \mathrm{ufp}\left(\mathrm{float}(|p|^T |q|)\right), \\ c_2 = n+2, \quad d_2 = 0, \quad e_2 = \frac{n}{2}, \quad f_2 = \mathrm{ufp}\left(\mathrm{float}(|z|^T |w|)\right) \end{aligned}$$

として, 定理 4.4 を用いればよい. $\ell = \max(c_1, c_2)$ と置き, 補題 4.3 より $\varphi = \mathrm{fl}(\ell\mathbf{u} + (4\ell + 27)\mathbf{u}^2)$ とする. よって,

$$\mathrm{fl}\left(\left|\mathrm{float}(p^T q) - \mathrm{float}(z^T w)\right|\right) > \mathrm{fl}(\varphi \cdot (f_1 + f_2) + (1 + 6\mathbf{u}) \cdot ((e_1 + e_2) + 1)\mathbf{u}_S)$$

という論理式が真ならば, $\mathrm{float}(p^T q)$ と $\mathrm{float}(z^T w)$ の大小関係は $p^T q$ と $z^T w$ の大小関係と等しい.

また, (4.2) に属する丸め誤差評価式として

$$\left|x^T y - \mathrm{float}(x^T y)\right| \le n\mathbf{u}|x|^T |y| + \frac{n}{2}\mathbf{u}_S$$

が知られている [8]. ここで, x, y にそれぞれ $|x|, |y|$ を代入したものも成立する. すなわち, 補題 4.3 の仮定は満たされる. よって, 補題 4.3 を適用すれば, $2n < \mathbf{u}^{-1}$ ならば

$$\left|x^T y - \mathrm{float}(x^T y)\right| \le (n\mathbf{u} + 2n^2\mathbf{u}^2)\mathrm{float}(|x|^T |y|) + n\mathbf{u}_S$$

が成立するため,

$$\begin{aligned} c_1 = m, \quad d_1 = 2m^2, \quad e_1 = m, \quad f_1 = \mathrm{float}(|p|^T |q|), \\ c_2 = n, \quad d_2 = 2n^2, \quad e_2 = n, \quad f_2 = \mathrm{float}(|z|^T |w|) \end{aligned}$$

として定理 4.4 を使用すればよい. 補題 4.3 を利用すれば, $\ell = \max(c_1, c_2)$, $\omega = \max(2m^2, 2n^2)$, $\varphi = \mathrm{fl}(\ell\mathbf{u} + (2\ell^2 + 4\ell + \omega + 27)\mathbf{u}^2)$ である. ただし, $\omega < \mathbf{u}^{-1}$ に注意する必要がある.

4.2.3 ホーナー法に対する応用例

ホーナー法により計算された多項式の値の比較を扱う．$a_i, x \in \mathbb{F},\ 0 \le i \le n$ とする．このとき, n 次多項式 $\sum_{i=0}^{n} a_i x^i$ に対するホーナー法では

$$a_0 + x(a_1 + x(a_2 + \cdots + x(a_{n-1} + xa_n)))$$

と変形して計算する．これに対して数値計算を用いた, n 次多項式 $\sum_{i=0}^{n} a_i x^i$ に対するホーナー法の表記を以下に定める．

$$\begin{aligned} H_k(a,x) &= \mathrm{fl}\left(H_{k+1}(a,x)\cdot x + a_k\right), \quad H_n(a,x) = a_n, \\ \tilde{H}_k(a,x) &= \mathrm{fl}\left(\tilde{H}_{k+1}(a,x)\cdot |x| + |a_k|\right), \quad \tilde{H}_n(a,x) = |a_n|. \end{aligned}$$

ここでは $k = n-1, n-2, \ldots, 0$ である．また, $H_0(a,x)$ が $\sum_{i=0}^{n} a_i x^i$ の近似値である．(4.2) の形式を満たす丸め誤差評価式として参考文献 [8] により, $n < \frac{1}{2}(\mathbf{u}^{-\frac{1}{2}} - 1)$ のとき

$$\left| H_0(a,x) - \sum_{i=0}^{n} a_i x^i \right| \le 2n\mathbf{u} \sum_{i=0}^{n} \left| a_i x^i \right|$$

が示されている．ただし演算中にアンダーフローが発生しないことを仮定する．ここで a, x にそれぞれの成分を非負の値にしたものも成立する．すなわち, 補題 4.1 の仮定は満たされる．よって, 補題 4.1 を適用すれば,

$$\left| H_0(a,x) - \sum_{i=0}^{n} a_i x^i \right| \le (2n\mathbf{u} + 8n^2\mathbf{u}^2)\tilde{H}_0(a,x)$$

と, (4.1) に対応した評価を得る．$b_i, y \in \mathbb{F},\ 0 \le i \le m$ としたとき, $H_k(a,x), \tilde{H}_k(a,x)$ と同様に $\sum_{i=0}^{m} b_i y^i$ に対して

$$\begin{aligned} I_k(b,y) &= \mathrm{fl}\left(I_{k+1}(b,y)\cdot y + b_k\right), \quad I_m(b,y) = b_m, \\ \tilde{I}_k(b,y) &= \mathrm{fl}\left(\tilde{I}_{k+1}(b,y)\cdot |y| + |b_k|\right), \quad \tilde{I}_m(b,y) = |b_m| \end{aligned}$$

を定義する．このとき, $\tilde{H}_0(a,x)$ と $\tilde{I}_0(b,y)$ の大小関係から真の関係を保証する浮動小数点フィルタは,

$$\begin{aligned} c_1 = 2n, \quad d_1 = 8n^2, \quad e_1 = 0, \quad f_1 = \tilde{H}_0(a,x), \\ c_2 = 2m, \quad d_2 = 8m^2, \quad e_2 = 0, \quad f_2 = \tilde{I}_0(b,y) \end{aligned}$$

として定理 4.4 を使用すればよい．補題 4.3 を利用すれば, $\ell = \max(c_1, c_2)$, $\omega = \max(8m^2, 8n^2)$, $\varphi = \mathrm{fl}(\ell\mathbf{u} + (2\ell^2 + 4\ell + \omega + 27)\mathbf{u}^2)$ である．ただし, $\omega < \mathbf{u}^{-1}$ に注意する必要がある．

(4.10) の評価を利用することで得られた浮動小数点フィルタ (4.11) において φ を $\varphi - 2\mathbf{u}$ にすることはできないことを示す．ただし, (4.10) の float の評価は fl とする．

Remark 6

長さ n (n は偶数, $4 \le n \le \mathbf{u}^{-1}/8$) のベクトル x, y を以下のように作成する.

$$x_i = \begin{cases} 1, & i = 1 \\ 3\mathbf{u}, & 2 \le i \le \dfrac{n}{2} \\ -\mathbf{u}, & \dfrac{n}{2} + 1 \le i \le n \end{cases}, \quad y_j = \begin{cases} 1, & j = 1 \\ \mathbf{u}, & 2 \le j \le n \end{cases}.$$

この x, y に対して成分の総和を計算する. ここで, $\mathrm{fl}(\sum_{i=1}^{n} x_i)$ は $\mathrm{fl}(x_1 + x_2)$ をはじめに計算し, その結果に x_3 を浮動小数点演算により加え, さらに x_4 を ... という順番で x_n まで浮動小数点演算により足し込んだ結果とする. $\mathrm{fl}(\sum_{j=1}^{n} y_j)$ も同様とする. ベクトル x と y の総和はそれぞれ

$$\sum_{i=1}^{n} x_i = 1 + (n-3)\mathbf{u}, \quad \sum_{i=1}^{n} y_i = 1 + (n-1)\mathbf{u}, \quad \sum_{i=1}^{n} x_i - \sum_{j=1}^{n} y_j = -2\mathbf{u} < 0$$

である. 一方で, $\mathrm{fl}(1 + \mathbf{u}) = 1$, $\mathrm{fl}(1 + 3\mathbf{u}) = 1 + 4\mathbf{u}$, $\mathrm{fl}((1 + 4k\mathbf{u}) + (-\mathbf{u})) = 1 + 4k\mathbf{u}$, ($k$: 自然数) であることから,

$$\mathrm{fl}\left(\sum_{i=1}^{n} x_i\right) = 1 + (2n-4)\mathbf{u}, \quad \mathrm{fl}\left(\sum_{j=1}^{n} y_j\right) = 1,$$

$$X := \mathrm{fl}\left(\mathrm{fl}\left(\sum_{i=1}^{n} x_i\right) - \mathrm{fl}\left(\sum_{j=1}^{n} y_j\right)\right) = (2n-4)\mathbf{u} > 0$$

となり, 浮動小数点演算により大小関係を判定した結果は, 真の大小関係と異なる. 次に総和に対する浮動小数点フィルタとして (4.11) の右辺 $\mathrm{fl}\left(\varphi \cdot (f_1 + f_2) + (1 + 6\mathbf{u})\mathbf{u}_S\right)$ の φ を $\varphi - 2\mathbf{u}$ で置き換えたものを考える. n は $4 \le n \le \mathbf{u}^{-1}/8$ を満たす偶数であるため, f_1 も f_2 も 1 となり $\mathrm{fl}(f_1 + f_2) = 2$ となる. ここで

$$\begin{aligned} &\mathrm{fl}((\varphi - 2\mathbf{u}) \cdot (f_1 + f_2) + (1 + 6\mathbf{u})\mathbf{u}_S) \\ =&\mathrm{fl}(\mathrm{fl}(\mathrm{fl}((n-1)\mathbf{u} + (4(n-1) + 27)\mathbf{u}^2) - 2\mathbf{u})2) \\ \le&(2n-5)\mathbf{u} \end{aligned}$$

であることから

$$X > \mathrm{fl}((\varphi - 2\mathbf{u}) \cdot (f_1 + f_2) + (1 + 6\mathbf{u})\mathbf{u}_S)$$

となり, φ を $\varphi - 2\mathbf{u}$ に変えた浮動小数点フィルタでは, 浮動小数点演算による大小関係の判定が正しいと判断してしまう. これは φ を $\varphi - 2\mathbf{u}$ に置き換えられないことを意味する.

5　誤差上限が既知であるときの評価

4 章では, 2 つの計算値の大小判定に対する誤差上限を考えたが, 本章ではその応用を考える. 例えば, 得られた誤差上限が符号に影響するほどなのかどうかや, 絶対誤差や相対誤差のチェックを行いたい場合もある. そこで, 数値計算結果と誤差評価が得られている場合に, 符号のチェック, 絶対誤差・相対誤差がある定数を超えていないかどうかをチェックする浮動小数点フィルタを提案する. また 4 章では, それぞれの誤差上限が浮動小数点演算により得られていることを前提としていた. そこで, 誤差評価が既知の 2 数の和および積を行った場合の誤差上限を求める方法も提案する. なお, 本章の内容は研究業績 [R3, R4, R11, R12, R20–R22] と関連がある.

5.1　前提条件

ある 2 つの計算式に対して, 真の値と数値計算結果による値をそれぞれ $a, b \in \mathbb{R}$, $x, y \in \mathbb{F}$ とする. これらに対して, 次の関係が成り立つことを仮定する.

$$|a - x| \le \left(c_1\mathbf{u} + d_1\mathbf{u}^2\right) f_1 + e_1\mathbf{u}_S, \tag{5.1}$$

$$|b - y| \le \left(c_2\mathbf{u} + d_2\mathbf{u}^2\right) f_2 + e_2\mathbf{u}_S. \tag{5.2}$$

ただし, $i = 1, 2$ に対して c_i, d_i, $e_i \in \mathbb{F}$ は 0 以上の整数, $0 \le f_i \in \mathbb{F}$ とする. また, これらは以下を満たすと仮定する.

$$c_i < \mathbf{u}^{-1}, \quad d_i < \mathbf{u}^{-1}.$$

(5.1), (5.2) に対して, 「符号の保証・絶対誤差の上限・相対誤差の上限・四則演算の誤差上限」を求める.

5.2　符号・絶対誤差・相対誤差の評価

浮動小数点フィルタは, 丸めモードの変更により簡単に作成することができる. 例えば, (5.1) の誤差評価に, $0 \le k \in \mathbb{F}$ に対して

$$k > \mathrm{fl}_{\triangle}((c_1\mathbf{u} + d_1\mathbf{u}^2)\cdot f + e_1\mathbf{u}_S)$$

が成立するのであれば, 絶対誤差が k 未満であることが保証される. 同様にして, $X = \mathrm{fl}_{\triangle}((c_1\mathbf{u} + d_1\mathbf{u}^2)\cdot f_1 + e_1\mathbf{u}_S)$ として

$$X < \mathrm{fl}_{\triangledown}(k\cdot(|x| - X))$$

が成立するのであれば, 相対誤差が k 未満であることが保証される. ただし, 丸めのモードを変更することができない環境や, 丸めのモードの変更は可能であるがコストがかかる環境

がある．そこで, 以降において, 丸めのモードを roundTiesToEven とし, 丸めのモード変更を利用しない浮動小数点フィルタを提案する．ここで, roundTiesToEven は多くの計算機環境のデフォルトであるため, 提案する浮動小数点フィルタがそのまま適用できる．

5.2.1 アンダーフローが起きない場合

まず, 特別な場合を考える．浮動小数点演算中にアンダーフローが発生しない, すなわち (5.1) において $e_1 = 0$ とする．このとき, 補題 2.8 と補題 2.10 より, 以下がすぐに導かれる．

定理 5.1 [R4]
$e_1 = 0$ とした (5.1) が与えられているとする．このとき

$$|x| > \mathrm{fl}((c_1\mathbf{u} + (d_1 + \mathrm{ufp}(c_1))\mathbf{u}^2)\cdot f_1)$$

が満たされれば, a と x の符号が同じであることが保証される．加えて, $0 < k \in \mathbb{F}$ に対して

$$k > \mathrm{fl}((c_1\mathbf{u} + (d_1 + \mathrm{ufp}(c_1))\mathbf{u}^2)\cdot f_1)$$

が満たされれば, x の絶対誤差が k 未満であることが保証される．

次に相対誤差についての浮動小数点フィルタを示す．

定理 5.2 [R4]
$e_1 = 0$ とした (5.1) が与えられているとする．$X = \mathrm{fl}((c_1\mathbf{u} + (d_1 + 3\mathrm{ufp}(c_1))\mathbf{u}^2)\cdot f_1)$, $Y = \mathrm{fl}(|x| - X)$, $r = \max\{2\mathbf{u}\cdot\mathrm{ufp}(Y), \mathbf{u}_S\}$ とする．このとき

$$X < \mathrm{fl}(k\cdot(Y - r)) \tag{5.3}$$

が満たされれば, 相対誤差が k 未満であることが保証される．

5.2.2 アンダーフローが起きうる場合

浮動小数点演算においてアンダーフローが発生し得る場合を考える．

定理 5.3 [R4]
(5.1) が与えられているとする．ここで

$$|x| > \mathrm{fl}((c_1\mathbf{u} + (d_1 + 3\mathrm{ufp}(c_1))\mathbf{u}^2)\cdot f_1 + (e_1 + 1)\mathbf{u}_S)$$

が満たされれば, a と x の符号が同じであることが保証される．

定理 5.4 [R4]
(5.1) が与えられているとする．ここで, $0 < k \in \mathbb{F}$ に対して

$$k > \mathrm{fl}((c_1\mathbf{u} + (d_1 + 3\mathrm{ufp}(c_1))\mathbf{u}^2)\cdot f_1 + (e_1 + 1)\mathbf{u}_S)$$

が満たされれば, x の絶対誤差が k 未満であることが保証される．

定理 5.3 と同様にして示せる.

次に相対誤差についての浮動小数点フィルタを示す.

定理 5.5 [R4]
(5.1) を仮定し, $X = \mathrm{fl}((c_1\mathbf{u} + (d_1 + 3\mathrm{ufp}(c_1))\mathbf{u}^2)\cdot f_1)$, $Y = \mathrm{fl}(X + (e_1 + 1)\mathbf{u}_S)$, $Z = \mathrm{fl}(|x| - Y)$, $r = \max\{2\mathbf{u}\cdot\mathrm{ufp}(Z), \mathbf{u}_S\}$ とする. このとき

$$Y < \mathrm{fl}(k\cdot(Z - r)) \tag{5.4}$$

が満たされれば, 相対誤差が $0 \le k \in \mathbb{F}$ 未満であることが保証される.

5.2.3 FMA を利用した評価

ここでは, FMA を利用した評価を提案する. 次の定理は, 補題 2.8 と補題 2.10 により直ちに導かれる.

定理 5.6 [R4]
(5.1) が与えられているとする. ここで

$$|x| > \mathrm{FMA}(\mathrm{fl}(c_1\mathbf{u} + (d_1 + \mathrm{ufp}(c_1))\mathbf{u}^2), f_1, \mathrm{fl}((e_1 + 1)\mathbf{u}_S))$$

が満たされれば, a と x の符号が同じであることが保証される. 加えて, $0 < k \in \mathbb{F}$ に対して

$$k > \mathrm{FMA}(\mathrm{fl}(c_1\mathbf{u} + (d_1 + \mathrm{ufp}(c_1))\mathbf{u}^2), f_1, \mathrm{fl}((e_1 + 1)\mathbf{u}_S))$$

が満たされれば, x の絶対誤差が k 未満であることが保証される.

次に相対誤差についての浮動小数点フィルタを示す.

定理 5.7 [R4]
(5.1) を仮定し, $X = \mathrm{FMA}(\mathrm{fl}(c_1\mathbf{u} + (d_1 + \mathrm{ufp}(c_1))\mathbf{u}^2), f_1, \mathrm{fl}((e_1 + 1)\mathbf{u}_S))$, $Y = \mathrm{fl}(|x| - X)$, $r = \max\{2\mathbf{u}\cdot\mathrm{ufp}(Y), \mathbf{u}_S\}$ とする. このとき

$$|x| > \mathrm{fl}(k\cdot(Y - r)) \tag{5.5}$$

が満たされれば, 相対誤差が $0 \le k \in \mathbb{F}$ 未満であることが保証される.

定理 5.6 および定理 5.7 はアンダーフローが発生することを仮定した. アンダーフローが発生しない場合には, それぞれ $\mathbf{u}_S$ の項を 0 として利用すればよい.

5.3 和・積の誤差上限

ここで, $c, d, \alpha \in \mathbb{F}$ を以下のように定義する.

$$j := \arg\max_i \left(c_i\mathbf{u} + d_i\mathbf{u}^2\right),$$

$$\alpha := c_j\mathbf{u} + d_j\mathbf{u}^2,$$
$$c := c_j, \quad d := d_j.$$

誤差評価が既知であるときに, 和および積の演算後の誤差上限を求める. ただし, 求める誤差上限は以下の形式とする.

$$\left\{\mathrm{fl}(\ldots)\mathbf{u} + \mathrm{fl}(\ldots)\mathbf{u}^2\right\}\mathrm{fl}\,(\ldots) + \mathrm{fl}\,(\ldots)\,\mathbf{u}_S.$$

得られる誤差上限は (5.1), (5.2) の形式を満たしている. すなわち, 得られた誤差上限をさらに用いて別の誤差上限を求めることもできる.

5.3.1 和および差

定理 5.8 [R3]
(5.1),(5.2) が得られている場合に, $\mathrm{fl}(x \pm y)$, $a \pm b$ に対する誤差上限は以下のようになる. ただし, $\pm$ は複号同順とする.

$$|\mathrm{fl}(x \pm y) - (a \pm b)| \leq \left(c_3\mathbf{u} + d_3\mathbf{u}^2\right) f_3 + e_3\mathbf{u}_S.$$

ただし,

$$\begin{aligned} c_3 &= \mathrm{fl}(c+1), \\ d_3 &= \mathrm{float}\,(2c+d+9), \\ f_3 &= \mathrm{fl}(f_1+f_2), \\ e_3 &= \lceil \mathrm{fl}\,((1+2\mathbf{u})\cdot(e_1+e_2))\rceil \end{aligned}$$

である.

条件 $e_i < \mathbf{u}^{-1}$, $(i = 1, 2)$ をつけ加えるのであれば, e_3 は $\lceil \mathrm{fl}\,((e_1+e_2)+2)\rceil$ と評価できる.

5.3.2 積

定理 5.9 [R3]
(5.1),(5.2) が得られており, 次の条件を満たしていると仮定する.

$$|x| \leq f_1, \quad |y| \leq f_2.$$

このとき, $\mathrm{fl}(x\cdot y)$, ab に対する誤差上限は以下のようになる.

$$|\mathrm{fl}(x\cdot y) - ab| \leq \left(c_3\mathbf{u} + d_3\mathbf{u}^2\right) f_3 + e_3\mathbf{u}_S.$$

ただし,

$$\begin{aligned} c_3 &= \mathrm{fl}\,(c_1 + c_2 + 1), \\ d_3 &= \mathrm{float}\,(10c_1 + 10c_2 + 2d_1 + 2d_2 + c_1\cdot c_2 + 67), \end{aligned}$$

$$
\begin{aligned}
f_3 &= \mathrm{fl}\left(f_1 \cdot f_2\right), \\
e_3 &= \lceil \mathrm{float}\left((2+12\mathbf{u})\cdot(f_1\cdot e_2) + (2+12\mathbf{u})\cdot(f_2\cdot e_1) + 3 + (1+6\mathbf{u})\cdot((e_1\cdot e_2)\mathbf{u}_S)\right) \rceil
\end{aligned}
$$

である.

次に, 誤差上限に関数 ufp を利用している場合を考える. このとき, 定理 5.9 の仮定を満たさない場合がある. そこで x, y に対する条件を ufp の性質を満たすように緩和し, 誤差解析を同様に行うことで導ける.

定理 5.10 [R3]
(5.1),(5.2) が得られており, 次の条件を満たしていると仮定する.

$$
\begin{aligned}
f_1 &= \mathrm{ufp}(f_1), \quad |x| \le 2f_1, \\
f_2 &= \mathrm{ufp}(f_2), \quad |y| \le 2f_2.
\end{aligned}
$$

このとき, $\mathrm{fl}(x\cdot y)$, ab に対する誤差上限は以下のようになる.

$$
|\mathrm{fl}(x\cdot y) - ab| \le \left(c_3\mathbf{u} + d_3\mathbf{u}^2\right) f_3 + e_3\mathbf{u}_S.
$$

ただし,

$$
\begin{aligned}
c_3 &= \mathrm{fl}\left(2c_1 + 2c_2 + 4\right), \\
d_3 &= \mathrm{float}\left(10c_1 + 10c_2 + 3d_1 + 3d_2 + c_1\cdot c_2 + 97\right), \\
f_3 &= \mathrm{fl}\left(f_1 \cdot f_2\right), \\
e_3 &= \lceil \mathrm{float}\left((3+16\mathbf{u})\cdot(f_1\cdot e_2) + (3+16\mathbf{u})\cdot(f_2\cdot e_1) + 4 + (1+6\mathbf{u})\cdot((e_1\cdot e_2)\mathbf{u}_S)\right) \rceil
\end{aligned}
$$

ただし, 定理 5.10 は [R3] で提案したものを改善したものである.

6　結論

6.1　本論文の結論

数値計算の結果の信頼性を評価する方法は, 精度保証付き数値計算の分野で研究されている. 例えば, ベクトルの総和や行列積など, 様々な誤差評価が提案されている. しかし, 誤差評価が提案されていない場合には, 個別の問題に対して誤差解析をする必要がある. その例として, 実数入力を考慮した Orient2D に対する浮動小数点フィルタを 3 章, 2 つの計算値の大小判定に対する浮動小数点フィルタを 4 章において提案した. これらの誤差解析においては, 過大評価をなるべく抑えるために, 非常に煩雑な丸め誤差解析となっていた. このような丸め誤差解析を個別の問題に対して行うのは, 体力的にも精神的にも大変である. ただし, 誤差評価を行いたい式の一部は, 既存の誤差評価を適用できる場合もある. そこで 5 章において, 誤差上限が既知であるときの評価を 2 種類提案した. 1 つは, 誤差上限が与えられた状態から, 符号・絶対誤差・相対誤差について評価する方法である. これにより, 誤差上限が得られた場合には, さらに丸め誤差解析を続けることなく評価することが可能となる. もう 1 つは, 誤差上限が与えられた状態から, 和および積を行った場合の誤差上限を求める方法である. これにより, ベクトルの総和と多項式どうしの和のように, 種類の異なる誤差上限どうしの演算に対しても評価をすることが可能となる.

以上のように, 従来の丸め誤差解析を拡張する方法を示した. 種類の異なる誤差上限どうしの演算の丸め誤差解析を比較的容易にすることで, 精度保証付き数値計算に対する貢献ができると考える.

6.2　課題と展望

本論文では, 誤差上限が既知である状態から, 和および積を行った場合の誤差上限を求める方法を提案した. 商と根号についての誤差上限を評価することが今後の課題である. これは, IEEE 754 が要請している基本演算が和・差・積・商・根号の 5 つだからである. 誤差上限が既知であるときの和・差・積・商・根号の誤差上限を求める方法が提案できれば, 数式処理システムに適用することで, 丸め誤差解析を簡略化・自動化することができる. これにより, 丸め誤差解析は難しくて行えない場合でも, 数式処理システムに入力するだけで, 誤差上限を評価することができるようになる. これにより, 数値計算の誤差評価を適用できる分野が増え, さらに精度保証付き数値計算に対する貢献ができると考える.

付録 A　誤差解析

ここでは, 浮動小数点フィルタを作成する際に, 省略した定理等の証明・誤差評価について記述する.

それぞれの証明の丸め誤差解析において, 等式および不等式の評価について, 手計算で求めたものを記述している. その後, 等式については差をとると同じになること, 不等式については差をとると正になることを, 数式処理システム (MATLAB の Symbolic Math Toolbox にあるシンボリック計算) を利用して成立することを確認している. 特に, 5 章の定理の証明 (A.4) で確認した.

A.1　2 章の定理の証明

補題 2.6 の証明

(2.6) より, 次のようにして導ける.

$$\begin{aligned}
\sum_{i=1}^{n} p_i &\le \mathrm{float}\left(\sum_{i=1}^{n} p_i\right) + (n-1)\mathbf{u}\cdot\mathrm{ufp}\left(\mathrm{float}\left(\sum_{i=1}^{n}|p_i|\right)\right) \\
&\le \mathrm{float}\left(\sum_{i=1}^{n} p_i\right) + (n-1)\mathbf{u}\cdot\mathrm{float}\left(\sum_{i=1}^{n}|p_i|\right) \\
&\le \mathrm{float}\left(\sum_{i=1}^{n} |p_i|\right) + (n-1)\mathbf{u}\cdot\mathrm{float}\left(\sum_{i=1}^{n}|p_i|\right) \\
&= (1+(n-1)\mathbf{u})\cdot\mathrm{float}\left(\sum_{i=1}^{n}|p_i|\right) \\
&\le (1+\mathbf{u})^{n-1}\,\mathrm{float}\left(\sum_{i=1}^{n}|p_i|\right)
\end{aligned}$$

□

補題 2.7 の証明

$a > b$ より, a の次に小さい浮動小数点数は b 以上である. ここで $a \le \mathbf{u}_N$ か否か, $a > \mathbf{u}_N$ のとき a が 2 のべき乗数か否かで場合分けして考えると,

$$\begin{cases}
a - 2\mathbf{u}\cdot\mathrm{ufp}(a) \ge b & a \in \mathbb{F}\backslash\mathbb{U},\ a \ne \mathrm{ufp}(a), \\
a - \mathbf{u}\cdot\mathrm{ufp}(a) \ge b & a \in \mathbb{F}\backslash\mathbb{U},\ a = \mathrm{ufp}(a),\ a \ne \mathbf{u}_N, \\
a - \mathbf{u}_S \ge b & a \in \mathbb{U}\cup\{\mathbf{u}_N\}
\end{cases}$$

が成立する. $a \in \mathbb{U}\cup\{\mathbf{u}_N\}$ のときは $\mathbf{u}\cdot\mathrm{ufp}(a) < \mathbf{u}_S$ より $a \ge b + \mathbf{u}\cdot\mathrm{ufp}(a)$ が成り立つ. これらより命題は成立する. □

補題 2.9 の証明

2 つに場合分けして考える.

- $\mathrm{fl}(a\cdot b) \geq \mathbf{u}_N$ の場合

  $X := \max\{2\mathbf{u}\cdot\mathrm{ufp}(a), \mathbf{u}_S\}$ とする. 定理 2.2, (2.7), (2.9) より以下を得る.

$$\begin{aligned}\mathrm{fl}(\mathrm{succ}(a)\cdot b) &= \mathrm{fl}((a+X)\cdot b)\\ &\geq (a+X)b - \mathbf{u}(a+X)b\\ &= ab + ((1-\mathbf{u})X - a\mathbf{u})b\\ &\geq ab + ((1-\mathbf{u})X - (2\mathrm{ufp}(a) - X)\mathbf{u})b\\ &= ab + (X - 2\mathbf{u}\cdot\mathrm{ufp}(a))b\\ &= ab + (\max\{2\mathbf{u}\cdot\mathrm{ufp}(a), \mathbf{u}_S\} - 2\mathbf{u}\cdot\mathrm{ufp}(a))b \geq ab\end{aligned}$$

  よって $ab \leq \mathrm{fl}(\mathrm{succ}(a)\cdot b)$ が成り立つ.

- $\mathrm{fl}(a\cdot b) < \mathbf{u}_N$ の場合

  仮定と定理 2.2 から以下が成り立つ.

$$ab \leq \mathrm{fl}(a\cdot b) + \frac{\mathbf{u}_S}{2} \leq \mathrm{fl}(\mathrm{succ}(a)\cdot b) + \frac{\mathbf{u}_S}{2}$$

以上より, (2.11) が成立する. □

補題 2.10 の証明

$c = 0$ または $d = 0$ のとき, (2.12) は自明に満たされる. よって, $c \neq 0$ かつ $d \neq 0$ について議論すればよい. もし $\mathrm{fl}(c\mathbf{u} + d\mathbf{u}^2) = c\mathbf{u} + d\mathbf{u}^2$ であれば, (2.12) は満たされる. なぜならば, $c\mathbf{u} = \mathrm{fl}(c\mathbf{u})$, $d\mathbf{u}^2 = \mathrm{fl}(d\mathbf{u}^2)$, $d < \mathrm{fl}(d + \mathrm{ufp}(c))$ が成り立つからである. 以後, $\mathrm{fl}(c\mathbf{u} + d\mathbf{u}^2) \neq c\mathbf{u} + d\mathbf{u}^2$ の場合についてのみ考える.

$\mathrm{fl}(c\mathbf{u} + d\mathbf{u}^2) \neq c\mathbf{u} + d\mathbf{u}^2$ であるから, $c\mathbf{u} + d\mathbf{u}^2 \notin \mathbb{F}$ が言える. すなわち, $c\mathbf{u} + d\mathbf{u}^2$ は隣り合う 2 つの浮動小数点数の間に存在する. つまり, ある定数 $k \in \mathbb{N}_0$ が存在して, 以下を満たす.

$$\alpha_1 := c\mathbf{u} + 2k\mathrm{ufp}(c)\mathbf{u}^2 < c\mathbf{u} + d\mathbf{u}^2 < c\mathbf{u} + 2(k+1)\mathrm{ufp}(c)\mathbf{u}^2 =: \alpha_2$$

ただし, $\alpha_2 = \mathrm{succ}(\alpha_1)$ である. また, $\mathrm{ufp}(c\mathbf{u} + d\mathbf{u}^2) = \mathrm{ufp}(c\mathbf{u})$ となるため,

$$|c\mathbf{u} + (d + \mathrm{ufp}(c))\mathbf{u}^2 - \alpha_1| > |c\mathbf{u} + (d + \mathrm{ufp}(c))\mathbf{u}^2 - \alpha_2|$$

が成り立つ. すなわち, $\mathrm{fl}(c\mathbf{u} + (d + \mathrm{ufp}(c))\mathbf{u}^2) = \alpha_2$ となる. 以上より, (2.12) が成り立つ. □

補題 2.11 の証明

補題 2.10 より

$$(c\mathbf{u} + d\mathbf{u}^2)f \leq \mathrm{fl}(c\mathbf{u} + (d + \mathrm{ufp}(c))\mathbf{u}^2)f$$

が成り立つ.

ここで, 2 つに場合分けして考える.

- $\mathrm{ufp}(\mathrm{fl}(c\mathbf{u} + (d + \mathrm{ufp}(c))\mathbf{u}^2)) = \mathrm{ufp}(\mathrm{fl}(c\mathbf{u} + d\mathbf{u}^2))$ の場合
  仮定と (2.8) より, 以下を得る.

$$\begin{aligned}\mathrm{succ}(\mathrm{fl}(c\mathbf{u} + (d + \mathrm{ufp}(c))\mathbf{u}^2)) &= \mathrm{fl}(c\mathbf{u} + (d + \mathrm{ufp}(c))\mathbf{u}^2) + 2\mathrm{ufp}(c)\mathbf{u}^2 \\ &= \mathrm{fl}(c\mathbf{u} + (d + 3\mathrm{ufp}(c))\mathbf{u}^2). \qquad \text{(A.1.1)}\end{aligned}$$

  (A.1.1) を (2.11) に代入することで, (2.13) が満たされる.
- $\mathrm{ufp}(\mathrm{fl}(c\mathbf{u} + (d + \mathrm{ufp}(c))\mathbf{u}^2)) > \mathrm{ufp}(\mathrm{fl}(c\mathbf{u} + d\mathbf{u}^2))$ の場合
  このとき, ある 2 の冪乗数 $g \in \mathbb{N}$ が存在し, 以下を満たす.

$$\mathrm{ufp}(\mathrm{fl}(c\mathbf{u} + (d + \mathrm{ufp}(c))\mathbf{u}^2)) \geq g > \mathrm{ufp}(\mathrm{fl}(c\mathbf{u} + d\mathbf{u}^2)) \geq \mathrm{ufp}(c\mathbf{u} + d\mathbf{u}^2)$$

  よって, $(c\mathbf{u} + d\mathbf{u}^2)f$ は次のように評価できる.

$$\begin{aligned}(c\mathbf{u} + d\mathbf{u}^2)f \leq gf &\leq \mathrm{fl}(g \cdot f) + \frac{\mathbf{u}_S}{2} \\ &\leq \mathrm{fl}(\mathrm{ufp}(\mathrm{fl}(c\mathbf{u} + (d + \mathrm{ufp}(c))\mathbf{u}^2)) \cdot f) + \frac{\mathbf{u}_S}{2} \\ &\leq \mathrm{fl}((c\mathbf{u} + (d + \mathrm{ufp}(c))\mathbf{u}^2) \cdot f) + \frac{\mathbf{u}_S}{2}\end{aligned}$$

以上より, (2.13) が成り立つ.

□

補題 2.12 の証明

(2.14) について示す. ここで, 組み合わせの数を求める関数 $C(n, r)$ を定義する.

$$C(n, r) = \frac{n!}{r!(n - r)!}$$

$C(n, r)$ を利用すれば, $(1 + \mathbf{u})^k \left(c\mathbf{u} + d\mathbf{u}^2\right)$ を以下のように展開できる.

$$\begin{aligned}&(1 + \mathbf{u})^k \left(c\mathbf{u} + d\mathbf{u}^2\right) \\ =&c\mathbf{u} + \left(C(k, 1)c + C(k, 0)d\right)\mathbf{u}^2 + \left(C(k, 2)c + C(k, 1)d\right)\mathbf{u}^3 + \cdots \\ =&c\mathbf{u} + \frac{(1 + \mathbf{u})^3}{(1 + \mathbf{u})^3}\left(\left(C(k, 1)c + C(k, 0)d\right)\mathbf{u}^2 + \left(C(k, 2)c + C(k, 1)d\right)\mathbf{u}^3 + \cdots\right) \\ =&c\mathbf{u} + \frac{1}{(1 + \mathbf{u})^3} \\ &\left(\left(C(k, 1)c + C(k, 0)d\right)\mathbf{u}^2 + \left(3C(k, 1)c + C(k, 2)c + 3C(k, 0)d + C(k, 1)d\right)\mathbf{u}^3 + \cdots\right) \\ <&c\mathbf{u} + \frac{1}{(1 + \mathbf{u})^3} \\ &\left(\left(C(k, 1)c + C(k, 0)d + \left(3C(k, 1)c + C(k, 2)c + 3C(k, 0)d + C(k, 1)d + 1\right)\right)\mathbf{u}^2\right)\end{aligned}$$

$$
\begin{aligned}
&= c\mathbf{u} + \frac{1}{(1+\mathbf{u})^3}\left(kc + d + 4 + 4k + \frac{k(k-1)}{2}\right)\mathbf{u}^2 \\
&= c\mathbf{u} + \frac{1}{(1+\mathbf{u})^3}\left(kc + d + const\right)\mathbf{u}^2
\end{aligned}
$$

4 行目から 5 行目の式変形は, $\mathbf{u}^{i+1}$ の項から $\mathbf{u}^i$ の項への繰り上がりは最大で $\mathbf{u}^i$ の係数 $+1$ であることを利用している. ここで, $kc + d + const < \mathbf{u}^{-1}$ として k について解くと

$$
0 \leq k \leq \left\lfloor \frac{(1-\sqrt{2})(2c+7) + 2\sqrt{2(\mathbf{u}^{-1} - d - 4)}}{2\sqrt{2}} \right\rfloor
$$

となる.

(2.15), (2.16) は (2.14) に対して定理 2.2 を適用し, その上限を評価することで導けるため, 証明は省略する. □

A.2 3 章の定理の証明

定理 3.1 の証明

入力値を丸めた値 a_x, a_y, b_x, b_y, c_x, c_y が正規化数・零・非正規化数のいずれかである場合, 入力値に対する行列式 $\det(G')$ は (3.1) と (3.2) より次のように式変形できる.

$$
\begin{aligned}
\det(G') = \begin{vmatrix} a'_x & a'_y & 1 \\ b'_x & b'_y & 1 \\ c'_x & c'_y & 1 \end{vmatrix} &= \begin{vmatrix} a_x + r_{ax} + \eta_1 & a_y + r_{ay} + \eta_2 & 1 \\ b_x + r_{bx} + \eta_3 & b_y + r_{by} + \eta_4 & 1 \\ c_x + r_{cx} + \eta_5 & c_y + r_{cy} + \eta_6 & 1 \end{vmatrix} \\
&= (p_5 - p_6) + K_1 + K_2
\end{aligned}
$$

ただし, K_1, K_2 は以下である.

$$
\begin{aligned}
K_1 = \quad & (r_{ax} - r_{cx})(b_y - c_y) - (a_y - c_y)(r_{bx} - r_{cx}) \\
& + (a_x - c_x)(r_{by} - r_{cy}) - (r_{ay} - r_{cy})(b_x - c_x) \\
& + (r_{ax} - r_{cx})(r_{by} - r_{cy}) - (r_{ay} - r_{cy})(r_{bx} - r_{cx}), \\
K_2 = \quad & (\eta_1 - \eta_5)(b_y - c_y) - (a_y - c_y)(\eta_3 - \eta_5) \\
& + (a_x - c_x)(\eta_4 - \eta_6) - (\eta_2 - \eta_6)(b_x - c_x) \\
& + (\eta_1 - \eta_5)(r_{by} - r_{cy}) - (r_{ay} - r_{cy})(\eta_3 - \eta_5) \\
& + (r_{ax} - r_{cx})(\eta_4 - \eta_6) - (\eta_2 - \eta_6)(r_{bx} - r_{cx}) \\
& + (\eta_1 - \eta_5)(\eta_4 - \eta_6) - (\eta_2 - \eta_6)(\eta_3 - \eta_5)
\end{aligned}
$$

よって, (A.2.3) は (3.3) より次のように評価される.

$$
\det(G') = (1+\delta_{13})\left(\mathrm{fl}(q_5 - q_6) + \frac{(M_3 - 1)\, q_5 - (M_4 - 1)\, q_6 + M_1 \cdot \eta_7 - M_2 \cdot \eta_8 + K_1 + K_2}{1 + \delta_{13}}\right)
$$

ここで, $\mathrm{fl}(q_5 - q_6)$ と (A.2.3) の符号が同じである判定条件を与える. 上式において $(1+\delta_{13})$ 倍することは符号に影響しないため, 次の不等式が成り立てば結果の符号が保証される.

$$
\mathrm{fl}\left(|q_5 - q_6|\right) > \left| \frac{(M_3 - 1)\, q_5 - (M_4 - 1)\, q_6 + M_1 \cdot \eta_7 - M_2 \cdot \eta_8 + K_1 + K_2}{1 + \delta_{13}} \right| \quad \text{(A.2.3)}
$$

ここで, K_1 の絶対値の上限は, (3.1) より以下のように評価できる.

$$\begin{aligned}
|K_1| \leq\ & (|r_{ax}|+|r_{cx}|)(|b_y|+|c_y|)+(|a_y|+|c_y|)(|r_{bx}|+|r_{cx}|)\\
&+(|a_x|+|c_x|)(|r_{by}|+|r_{cy}|)+(|r_{ay}|+|r_{cy}|)(|b_x|+|c_x|)\\
&+(|r_{ax}|+|r_{cx}|)(|r_{by}|+|r_{cy}|)+(|r_{ay}|+|r_{cy}|)(|r_{bx}|+|r_{cx}|)\\
\leq\ & (\mathbf{u}|a_x|+\mathbf{u}|c_x|)r_2+r_3(\mathbf{u}|b_x|+\mathbf{u}|c_x|)+r_1(\mathbf{u}|b_y|+\mathbf{u}|c_y|)+(\mathbf{u}|a_y|+\mathbf{u}|c_y|)r_4\\
&+(\mathbf{u}|a_x|+\mathbf{u}|c_x|)(\mathbf{u}|b_y|+\mathbf{u}|c_y|)+(\mathbf{u}|a_y|+\mathbf{u}|c_y|)(\mathbf{u}|b_x|+\mathbf{u}|c_x|)\\
=\ & \mathbf{u}r_1{\cdot}r_2+\mathbf{u}r_3{\cdot}r_4+\mathbf{u}r_1{\cdot}r_2+\mathbf{u}r_3{\cdot}r_4+\mathbf{u}^2r_1{\cdot}r_2+\mathbf{u}^2r_3{\cdot}r_4\\
=\ & (2\mathbf{u}+\mathbf{u}^2)(r_5+r_6)
\end{aligned}$$

同様に, K_2 の絶対値の上限は次のように評価できる.

$$|K_2| \leq \mathbf{u}_S(1+\mathbf{u})\left((r_1+r_2)+(r_3+r_4)\right)+2\mathbf{u}_S^2 \tag{A.2.7}$$

また, M_1, M_2, M_3-1, M_4-1 は次のように評価できる.

$$\begin{aligned}
&|M_1| \leq (1+\mathbf{u})^2, \quad |M_2| \leq (1+\mathbf{u})^2\\
&|M_3-1| \leq 3\mathbf{u}+3\mathbf{u}^2+\mathbf{u}^3, \quad |M_4-1| \leq 3\mathbf{u}+3\mathbf{u}^2+\mathbf{u}^3
\end{aligned}$$

これより, (A.2.3) の右辺の上限を定理 2.2, (3.4), (3.5), (A.2.4), (A.2.7), (A.2.8), (A.2.9) を用いて求める.

$$\begin{aligned}
&\text{不等式 } (A.2.3) \text{ の右辺}\\
\leq\ & \big((3\mathbf{u}+3\mathbf{u}^2+\mathbf{u}^3)|q_5|+(3\mathbf{u}+3\mathbf{u}^2+\mathbf{u}^3)|q_6|+\mathbf{u}_S(1+\mathbf{u})^2\\
&\quad +(2\mathbf{u}+\mathbf{u}^2)(r_5+r_6)+\mathbf{u}_S(1+\mathbf{u})\left((r_1+r_2)+(r_3+r_4)\right)+2\mathbf{u}_S^2\big)/(1-\mathbf{u})\\
=\ & \big((3\mathbf{u}+3\mathbf{u}^2+\mathbf{u}^3)(|q_5|+|q_6|)+(2\mathbf{u}+\mathbf{u}^2)(r_5+r_6)\\
&\quad +\mathbf{u}_S(1+\mathbf{u})\left((r_1+r_2)+(r_3+r_4)\right)+\mathbf{u}_S(1+2\mathbf{u}+\mathbf{u}^2+2\mathbf{u}_S)\big)/(1-\mathbf{u})\\
=\ & \left(3\mathbf{u}+6\mathbf{u}^2+7\mathbf{u}^3+\frac{7\mathbf{u}^4}{1-\mathbf{u}}\right)(|q_5|+|q_6|)+\left(2\mathbf{u}+3\mathbf{u}^2+\frac{3\mathbf{u}^3}{1-\mathbf{u}}\right)(r_5+r_6)\\
&\quad +\mathbf{u}_S\left(1+2\mathbf{u}+\frac{2\mathbf{u}^2}{1-\mathbf{u}}\right)\left((r_1+r_2)+(r_3+r_4)\right)\\
&\quad +\mathbf{u}_S\left(1+3\mathbf{u}+4\mathbf{u}^2+\frac{4\mathbf{u}^3+2\mathbf{u}_S}{1-\mathbf{u}}\right)\\
<\ & (3\mathbf{u}+6\mathbf{u}^2+8\mathbf{u}^3)(|q_5|+|q_6|)+(2\mathbf{u}+4\mathbf{u}^2)(r_5+r_6)\\
&\quad +\mathbf{u}_S(1+3\mathbf{u})\left((r_1+r_2)+(r_3+r_4)\right)+\mathbf{u}_S(1+3\mathbf{u}+5\mathbf{u}^2)\\
\leq\ & (3\mathbf{u}+6\mathbf{u}^2+8\mathbf{u}^3)(|q_5|+|q_6|)+(2\mathbf{u}+4\mathbf{u}^2)((1+\mathbf{u})^4\mathrm{fl}(s_5+s_6)+\mathbf{u}_S(1+\mathbf{u})^2)\\
&\quad +\mathbf{u}_S(1+3\mathbf{u})(1+\mathbf{u})^3\mathrm{fl}\left((s_1+s_2)+(s_3+s_4)\right)+\mathbf{u}_S(1+3\mathbf{u}+5\mathbf{u}^2)\\
\leq\ & (3\mathbf{u}+6\mathbf{u}^2+8\mathbf{u}^3)(1+\mathbf{u})\mathrm{fl}(|q_5|+|q_6|)+(2\mathbf{u}+4\mathbf{u}^2)(1+\mathbf{u})^4\mathrm{fl}(s_5+s_6)\\
&\quad +\mathbf{u}_S(1+3\mathbf{u})(1+\mathbf{u})^3\mathrm{fl}\left((s_1+s_2)+(s_3+s_4)\right)\\
&\quad +\mathbf{u}_S(1+5\mathbf{u}+13\mathbf{u}^2+10\mathbf{u}^3+4\mathbf{u}^4)\\
<\ & (3\mathbf{u}+9\mathbf{u}^2+15\mathbf{u}^3)\mathrm{fl}(|q_5|+|q_6|)+(2\mathbf{u}+12\mathbf{u}^2+29\mathbf{u}^3)\mathrm{fl}(s_5+s_6)\\
&\quad +\mathbf{u}_S(1+7\mathbf{u})\mathrm{fl}\left((s_1+s_2)+(s_3+s_4)\right)+\mathbf{u}_S(1+6\mathbf{u})
\end{aligned}$$

$=: L_1 + L_2 + L_3$

ただし, L_1, L_2, L_3 は以下である.

$$
\begin{aligned}
L_1 &= (3\mathbf{u} + 9\mathbf{u}^2 + 15\mathbf{u}^3)\mathrm{fl}(|q_5| + |q_6|) + (2\mathbf{u} + 12\mathbf{u}^2 + 29\mathbf{u}^3)\mathrm{fl}(s_5 + s_6) \\
L_2 &= \mathbf{u}_S(1 + 7\mathbf{u})\mathrm{fl}\left((s_1 + s_2) + (s_3 + s_4)\right) \\
L_3 &= \mathbf{u}_S(1 + 6\mathbf{u})
\end{aligned}
$$

これより, $L := L_1 + L_2 + L_3$ の上限を浮動小数点演算のみで評価する.

1. L の上限を評価した浮動小数点フィルタ

   L_1, L_2 の上限をとり, それらと L_3 の和の上限について考える. まず, L_1 の上限は, $\mathrm{fl}(|q_5| + |q_6|) = \mathrm{fl}(s_5 + s_6) = 0$ となりうることを考慮して, 定理 2.2 から次のようになる.

$$
\begin{aligned}
L_1 &\leq \frac{3\mathbf{u} + 21\mathbf{u}^2 + 70\mathbf{u}^3}{(1+\mathbf{u})^4}\mathrm{fl}\left(|q_5| + |q_6|\right) + \frac{2\mathbf{u} + 20\mathbf{u}^2 + 90\mathbf{u}^3}{(1+\mathbf{u})^4}\mathrm{fl}\left(s_5 + s_6\right) \\
&\leq \frac{\mathrm{fl}(3\mathbf{u} + 24\mathbf{u}^2)}{(1+\mathbf{u})^4}\mathrm{fl}\left(|q_5| + |q_6|\right) + \frac{\mathrm{fl}(2\mathbf{u} + 24\mathbf{u}^2)}{(1+\mathbf{u})^4}\mathrm{fl}\left(s_5 + s_6\right) \\
&\leq \frac{\mathrm{fl}((3\mathbf{u} + 24\mathbf{u}^2)\cdot(|q_5| + |q_6|))}{(1+\mathbf{u})^3} + \frac{\mathbf{u}_S}{2(1+\mathbf{u})^4} \\
&\quad + \frac{\mathrm{fl}((2\mathbf{u} + 24\mathbf{u}^2)\cdot(s_5 + s_6))}{(1+\mathbf{u})^3} + \frac{\mathbf{u}_S}{2(1+\mathbf{u})^4} \\
&\leq \frac{\mathrm{fl}((3\mathbf{u} + 24\mathbf{u}^2)\cdot(|q_5| + |q_6|) + (2\mathbf{u} + 24\mathbf{u}^2)\cdot(s_5 + s_6))}{(1+\mathbf{u})^2} + \frac{\mathbf{u}_S}{(1+\mathbf{u})^4}
\end{aligned}
$$

   $(A.2.13)$ から $(A.2.14)$ の式変形は, 浮動小数点数による上限の評価を表している. まず, $(A.2.13)$ の $3\mathbf{u}+21\mathbf{u}^2+70\mathbf{u}^3$ は浮動小数点数ではない. そこで, $3\mathbf{u}+21\mathbf{u}^2+70\mathbf{u}^3$ よりも大きく, 最も小さい浮動小数点数 $\mathrm{fl}(3\mathbf{u}+24\mathbf{u}^2)$ を上限としている. $2\mathbf{u}+20\mathbf{u}^2+90\mathbf{u}^3$ に対する $\mathrm{fl}(2\mathbf{u} + 24\mathbf{u}^2)$ も同様である. 以後の不等式の導出においても, ある実数よりも大きく, かつ最も小さい浮動小数点数を用いて証明を行う. L_1 の評価と同様にして, L_2 の上限は以下のように評価できる.

$$
\begin{aligned}
L_2 &\leq \frac{\mathbf{u}_S(1 + 11\mathbf{u})}{(1+\mathbf{u})^3}\mathrm{fl}\left((s_1 + s_2) + (s_3 + s_4)\right) \leq \frac{\mathrm{fl}(2\mathbf{u}_S)}{(1+\mathbf{u})^3}\mathrm{fl}\left((s_1 + s_2) + (s_3 + s_4)\right) \\
&\leq \frac{\mathrm{fl}\left(2\mathbf{u}_S((s_1 + s_2) + (s_3 + s_4))\right)}{(1+\mathbf{u})^2} + \frac{\mathbf{u}_S}{2(1+\mathbf{u})^3}
\end{aligned}
$$

   これと定理 2.2 より, L は次のように評価される.

$$
\begin{aligned}
L \leq & \frac{\mathrm{fl}((3\mathbf{u} + 24\mathbf{u}^2)\cdot(|q_5| + |q_6|) + (2\mathbf{u} + 24\mathbf{u}^2)\cdot(s_5 + s_6))}{(1+\mathbf{u})^2} + \frac{\mathbf{u}_S}{(1+\mathbf{u})^4} \\
& + \frac{\mathrm{fl}\left(2\mathbf{u}_S((s_1 + s_2) + (s_3 + s_4))\right)}{(1+\mathbf{u})^2} + \frac{\mathbf{u}_S}{2(1+\mathbf{u})^3} + \mathbf{u}_S(1 + 6\mathbf{u})
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{\mathrm{fl}((3\mathbf{u}+24\mathbf{u}^2)\cdot(|q_5|+|q_6|)+(2\mathbf{u}+24\mathbf{u}^2)\cdot(s_5+s_6))}{(1+\mathbf{u})^2} \\
&\quad + \frac{\mathrm{fl}(2\mathbf{u}_S((s_1+s_2)+(s_3+s_4)))}{(1+\mathbf{u})^2} \\
&\quad + \mathbf{u}_S\left(1+\frac{1}{(1+\mathbf{u})^4}+\frac{1}{2(1+\mathbf{u})^3}+6\mathbf{u}\right) \\
&< \frac{\mathrm{fl}((3\mathbf{u}+24\mathbf{u}^2)\cdot(|q_5|+|q_6|)+(2\mathbf{u}+24\mathbf{u}^2)\cdot(s_5+s_6)+2\mathbf{u}_S((s_1+s_2)+(s_3+s_4)))}{1+\mathbf{u}} \\
&\quad + \frac{\mathrm{fl}(3\mathbf{u}_S)}{1+\mathbf{u}} \\
&\le \mathrm{fl}\left((3\mathbf{u}+24\mathbf{u}^2)\cdot(|q_5|+|q_6|)+(2\mathbf{u}+24\mathbf{u}^2)\cdot(s_5+s_6)\right. \\
&\quad \left.+2\mathbf{u}_S((s_1+s_2)+(s_3+s_4))+3\mathbf{u}_S\right)
\end{aligned}
$$

CPU を用いた数値計算時に非正規化数が現れると計算速度が低下することがある. そこで $2\mathbf{u}_S, 3\mathbf{u}_S$ を $\mathbf{u}_N$ とおいて上限をとることにより, (3.6) が得られる.

2. L_1 の q_5, q_6 を絶対値で評価した浮動小数点フィルタ

   L_1 において, $|q_5| \le s_5$, $|q_6| \le s_6$ と評価できることを利用した場合は以下となる.

$$
\begin{aligned}
L_1 &\le (5\mathbf{u}+21\mathbf{u}^2+44\mathbf{u}^3)\mathrm{fl}\,(s_5+s_6) \le \frac{(5\mathbf{u}+36\mathbf{u}^2+123\mathbf{u}^3)\mathrm{fl}(s_5+s_6)}{(1+\mathbf{u})^3} \\
&\le \frac{\mathrm{fl}(5\mathbf{u}+40\mathbf{u}^2)\mathrm{fl}(s_5+s_6)}{(1+\mathbf{u})^3} \le \frac{\mathrm{fl}((5\mathbf{u}+40\mathbf{u}^2)\cdot(s_5+s_6))}{(1+\mathbf{u})^2} + \frac{\mathbf{u}_S}{2(1+\mathbf{u})^3}
\end{aligned}
$$

これより, L は次のように評価される.

$$
\begin{aligned}
L &\le \frac{\mathrm{fl}((5\mathbf{u}+40\mathbf{u}^2)\cdot(s_5+s_6))}{(1+\mathbf{u})^2} + \frac{\mathbf{u}_S}{2(1+\mathbf{u})^3} + \frac{\mathrm{fl}\,(2\mathbf{u}_S((s_1+s_2)+(s_3+s_4)))}{(1+\mathbf{u})^2} \\
&\quad + \frac{\mathbf{u}_S}{2(1+\mathbf{u})^3} + \mathbf{u}_S(1+6\mathbf{u}) \\
&\le \frac{\mathrm{fl}((5\mathbf{u}+40\mathbf{u}^2)\cdot(s_5+s_6)+2\mathbf{u}_S((s_1+s_2)+(s_3+s_4)))}{1+\mathbf{u}} \\
&\quad + \mathbf{u}_S\left(1+\frac{1}{(1+\mathbf{u})^3}+6\mathbf{u}\right) \\
&= \frac{\mathrm{fl}((5\mathbf{u}+40\mathbf{u}^2)\cdot(s_5+s_6)+2\mathbf{u}_S((s_1+s_2)+(s_3+s_4)))}{1+\mathbf{u}} \\
&\quad + \frac{\mathbf{u}_S}{1+\mathbf{u}}\left(1+\frac{1}{(1+\mathbf{u})^2}+7\mathbf{u}+6\mathbf{u}^2\right) \\
&< \frac{\mathrm{fl}((5\mathbf{u}+40\mathbf{u}^2)\cdot(s_5+s_6)+2\mathbf{u}_S((s_1+s_2)+(s_3+s_4)))}{1+\mathbf{u}} + \frac{\mathrm{fl}(3\mathbf{u}_S)}{1+\mathbf{u}} \\
&\le \mathrm{fl}\left((5\mathbf{u}+40\mathbf{u}^2)\cdot(s_5+s_6)+2\mathbf{u}_S((s_1+s_2)+(s_3+s_4))+3\mathbf{u}_S\right)
\end{aligned}
$$

同様に $2\mathbf{u}_S, 3\mathbf{u}_S$ を $\mathbf{u}_N$ とおいて上限をとれば, (3.7) が得られる.

□

定理 3.2 の証明

まず, (3.10), (3.11) について示す. 入力値を丸めた値 a_x, a_y, b_x, b_y, c_x, c_y が正規化数であるため (3.1) における $\eta_1, \ldots, \eta_6$ は 0 となり, 式 (A.2.3) において $K_2 = 0$ とできるため

$$\mathrm{fl}\left(|q_5 - q_6|\right) > \left| \frac{(M_3 - 1)\, q_5 - (M_4 - 1)\, q_6 + M_1 \cdot \eta_7 - M_2 \cdot \eta_8 + K_1}{1 + \delta_{13}} \right|$$

が成り立てば結果の符号が保証される. $K_2 = 0$ であるから (A.2.12) において L_2 を考慮する必要はなく, (A.2.15) の右辺の上限は $L_1 + L_3$ で評価できる. よって

不等式 $(A.2.15)$ の右辺
$< (3\mathbf{u} + 9\mathbf{u}^2 + 15\mathbf{u}^3)\mathrm{fl}(|q_5| + |q_6|) + (2\mathbf{u} + 12\mathbf{u}^2 + 29\mathbf{u}^3)\mathrm{fl}(s_5 + s_6) + \mathbf{u}_S(1 + 6\mathbf{u}) =: L_4$

となる. これより, L_4 の上限を浮動小数点演算のみで評価する.

1. L_4 の上限を評価した浮動小数点フィルタ

$$\begin{aligned}
L_4 &\leq \frac{3\mathbf{u} + 18\mathbf{u}^2 + 52\mathbf{u}^3}{(1+\mathbf{u})^3}\mathrm{fl}\left(|q_5| + |q_6|\right) + \frac{2\mathbf{u} + 18\mathbf{u}^2 + 72\mathbf{u}^3}{(1+\mathbf{u})^3}\mathrm{fl}\left(s_5 + s_6\right) \\
&\quad + \mathbf{u}_S(1 + 6\mathbf{u}) \\
&\leq \frac{\mathrm{fl}(3\mathbf{u} + 20\mathbf{u}^2)}{(1+\mathbf{u})^3}\mathrm{fl}\left(|q_5| + |q_6|\right) + \frac{\mathrm{fl}(2\mathbf{u} + 20\mathbf{u}^2)}{(1+\mathbf{u})^3}\mathrm{fl}\left(s_5 + s_6\right) + \mathbf{u}_S(1 + 6\mathbf{u}) \\
&\leq \frac{\mathrm{fl}((3\mathbf{u} + 20\mathbf{u}^2)\cdot(|q_5| + |q_6|))}{(1+\mathbf{u})^2} + \frac{\mathrm{fl}((2\mathbf{u} + 20\mathbf{u}^2)\cdot(s_5 + s_6))}{(1+\mathbf{u})^2} \\
&\quad + \frac{\mathbf{u}_S}{(1+\mathbf{u})^3} + \mathbf{u}_S(1 + 6\mathbf{u}) \\
&\leq \frac{\mathrm{fl}((3\mathbf{u} + 20\mathbf{u}^2)\cdot(|q_5| + |q_6|) + (2\mathbf{u} + 20\mathbf{u}^2)\cdot(s_5 + s_6))}{1+\mathbf{u}} \\
&\quad + \mathbf{u}_S\left(1 + \frac{1}{(1+\mathbf{u})^3} + 6\mathbf{u}\right) \\
&< \mathrm{fl}((3\mathbf{u} + 20\mathbf{u}^2)\cdot(|q_5| + |q_6|) + (2\mathbf{u} + 20\mathbf{u}^2)\cdot(s_5 + s_6) + 3\mathbf{u}_S)
\end{aligned}$$

$3\mathbf{u}_S$ を $\mathbf{u}_N$ とおいて上限をとれば, (3.10) が得られる.

2. L_4 の q_5, q_6 を絶対値で評価した浮動小数点フィルタ

L_4 において, $|q_5| \leq s_5$, $|q_6| \leq s_6$ と評価できることを利用した場合は以下となる.

$$\begin{aligned}
L_4 &\leq (5\mathbf{u} + 21\mathbf{u}^2 + 44\mathbf{u}^3)\mathrm{fl}\left(s_5 + s_6\right) + \mathbf{u}_S(1 + 6\mathbf{u}) \\
&\leq \frac{(5\mathbf{u} + 31\mathbf{u}^2 + 92\mathbf{u}^3)\mathrm{fl}(s_5 + s_6)}{(1+\mathbf{u})^2} + \mathbf{u}_S(1 + 6\mathbf{u}) \\
&\leq \frac{\mathrm{fl}((5\mathbf{u} + 32\mathbf{u}^2)\cdot(s_5 + s_6))}{1+\mathbf{u}} + \frac{\mathbf{u}_S}{2(1+\mathbf{u})^2} + \mathbf{u}_S(1 + 6\mathbf{u}) \\
&= \frac{\mathrm{fl}((5\mathbf{u} + 32\mathbf{u}^2)\cdot(s_5 + s_6))}{1+\mathbf{u}} + \mathbf{u}_S\left(1 + \frac{1}{2(1+\mathbf{u})^2} + 6\mathbf{u}\right)
\end{aligned}$$

$$
\begin{aligned}
&< \frac{\mathrm{fl}((5\mathbf{u}+32\mathbf{u}^2)\cdot(s_5+s_6))}{1+\mathbf{u}} + \frac{\mathrm{fl}(2\mathbf{u}_S)}{1+\mathbf{u}} \\
&\leq \mathrm{fl}((5\mathbf{u}+32\mathbf{u}^2)\cdot(s_5+s_6)+2\mathbf{u}_S)
\end{aligned}
$$

同様に $2\mathbf{u}_S$ を $\mathbf{u}_N$ とおいて上限をとれば, (3.11) が得られる.

次に, (3.8), (3.9) について示す.

入力値を丸めた値 a_x, a_y, b_x, b_y, c_x, c_y が正規化数であり, アンダーフローが発生しないため (A.2.15) において $\eta_7 = \eta_8 = 0$ とした

$$
\mathrm{fl}\left(|q_5 - q_6|\right) > \left| \frac{(M_3-1)\,q_5 - (M_4-1)\,q_6 + K_1}{1+\delta_{13}} \right|
$$

が成り立てば結果の符号が保証される. $K_2 = 0$ であるから $(A.2.12)$ において L_2 を考慮する必要はなく, また $\eta_7 = \eta_8 = 0$ であるため L_3 も考慮する必要はない. よって, (A.2.16) の右辺の上限は L_1 のみで評価できる. すなわち

$$
\begin{aligned}
&\text{不等式 } (A.2.16) \text{ の右辺} \\
&< (3\mathbf{u}+9\mathbf{u}^2+15\mathbf{u}^3)\mathrm{fl}(|q_5|+|q_6|) + (2\mathbf{u}+12\mathbf{u}^2+29\mathbf{u}^3)\mathrm{fl}(s_5+s_6) =: L_5
\end{aligned}
$$

となる. これより, L_5 の上限を浮動小数点演算のみで評価する. ただし, 仮定からアンダーフローを考慮しない.

1. L_5 の上限を評価した浮動小数点フィルタ

   L_5 の上限について考えると, $s_5 + s_6 \neq 0$ より次のようになる.

$$
\begin{aligned}
L_5 &< \frac{3\mathbf{u}+15\mathbf{u}^2+37\mathbf{u}^3}{(1+\mathbf{u})^2}\mathrm{fl}\left(|q_5|+|q_6|\right) + \frac{2\mathbf{u}+16\mathbf{u}^2+56\mathbf{u}^3}{(1+\mathbf{u})^2}\mathrm{fl}\left(s_5+s_6\right) \\
&< \frac{\mathrm{fl}(3\mathbf{u}+16\mathbf{u}^2)}{(1+\mathbf{u})^2}\mathrm{fl}\left(|q_5|+|q_6|\right) + \frac{\mathrm{fl}(2\mathbf{u}+20\mathbf{u}^2)}{(1+\mathbf{u})^2}\mathrm{fl}\left(s_5+s_6\right) \\
&\leq \frac{\mathrm{fl}((3\mathbf{u}+16\mathbf{u}^2)\cdot(|q_5|+|q_6|))}{(1+\mathbf{u})} + \frac{\mathrm{fl}((2\mathbf{u}+20\mathbf{u}^2)\cdot(s_5+s_6))}{(1+\mathbf{u})} \\
&\leq \mathrm{fl}((3\mathbf{u}+16\mathbf{u}^2)\cdot(|q_5|+|q_6|) + (2\mathbf{u}+20\mathbf{u}^2)\cdot(s_5+s_6))
\end{aligned}
$$

   これより, (3.8) が得られる.

2. L_5 の q_5, q_6 を絶対値で評価した浮動小数点フィルタ

   L_5 において, $|q_5| \leq s_5$, $|q_6| \leq s_6$ と評価できることを利用した場合は以下となる.

$$
\begin{aligned}
L_5 &\leq (5\mathbf{u}+21\mathbf{u}^2+44\mathbf{u}^3)\mathrm{fl}\left(s_5+s_6\right) < \frac{(5\mathbf{u}+26\mathbf{u}^2+66\mathbf{u}^3)\mathrm{fl}\left(s_5+s_6\right)}{1+\mathbf{u}} \\
&< \frac{\mathrm{fl}(5\mathbf{u}+32\mathbf{u}^2)\mathrm{fl}\left(s_5+s_6\right)}{1+\mathbf{u}} \leq \mathrm{fl}((5\mathbf{u}+32\mathbf{u}^2)\cdot(s_5+s_6))
\end{aligned}
$$

   これより, (3.9) が得られる.

□

A.3　4 章の定理の証明

補題 4.1 の証明

$(0 \leq)\ f' \leq f$ のときは自明であるため, $f' > f\ (\geq 0)$ のときを考える. 仮定より

$$f \geq f' - \left((c\mathbf{u} + d\mathbf{u}^2)f' + e\mathbf{u}_S\right) = \left(1 - (c\mathbf{u} + d\mathbf{u}^2)\right) f' - e\mathbf{u}_S$$

が成り立つ. これより, $f' \leq \dfrac{1}{1-(c\mathbf{u}+d\mathbf{u}^2)}(f + e\mathbf{u}_S)$ を得る. すなわち

$$\begin{aligned}(c\mathbf{u} + d\mathbf{u}^2)f' + e\mathbf{u}_S &\leq \frac{c\mathbf{u} + d\mathbf{u}^2}{1-(c\mathbf{u}+d\mathbf{u}^2)}(f + e\mathbf{u}_S) + e\mathbf{u}_S \\ &= \frac{c\mathbf{u} + d\mathbf{u}^2}{1-(c\mathbf{u}+d\mathbf{u}^2)} f + \frac{1}{1-(c\mathbf{u}+d\mathbf{u}^2)} e\mathbf{u}_S\end{aligned}$$

と上限をとれる. ここで, $x \in \mathbb{R}$ に対して $0 \leq x < \frac{1}{2}$ であるとき, 以下が成り立つ.

$$\frac{x}{1-x} \leq x + 2x^2, \quad \frac{1}{1-x} < 2$$

また, 仮定 $c\mathbf{u} + d\mathbf{u}^2 < \frac{1}{2}$ より $c\mathbf{u} < \frac{1}{2} - d\mathbf{u}^2$ と評価できる. これらを利用すると次のように評価できる.

$$\begin{aligned}&\frac{c\mathbf{u} + d\mathbf{u}^2}{1-(c\mathbf{u}+d\mathbf{u}^2)} f + \frac{1}{1-(c\mathbf{u}+d\mathbf{u}^2)} e\mathbf{u}_S \\ &< \left((c\mathbf{u} + d\mathbf{u}^2) + 2(c^2\mathbf{u}^2 + 2cd\mathbf{u}^3 + d^2\mathbf{u}^4)\right) f + 2e\mathbf{u}_S \\ &= \left(c\mathbf{u} + (2c^2+d)\mathbf{u}^2 + 4cd\mathbf{u}^3 + 2d^2\mathbf{u}^4\right) f + 2e\mathbf{u}_S \\ &\leq \left(c\mathbf{u} + (2c^2+d)\mathbf{u}^2 + 4\left(\frac{1}{2} - d\mathbf{u}^2\right)d\mathbf{u}^2 + 2d^2\mathbf{u}^4\right) f + 2e\mathbf{u}_S \\ &= \left(c\mathbf{u} + (2c^2+3d)\mathbf{u}^2 - 2d^2\mathbf{u}^4\right) f + 2e\mathbf{u}_S \leq \left(c\mathbf{u} + (2c^2+3d)\mathbf{u}^2\right) f + 2e\mathbf{u}_S\end{aligned}$$

□

補題 4.2 の証明

(2.4), (4.3) より, それぞれ次が成り立つ.

$$f_1 + f_2 \leq (1+\mathbf{u})\mathrm{fl}(f_1 + f_2), \quad |s_1 - s_2| \leq |s_1| + |s_2| \leq (e_1 + e_2)\mathbf{u}_S \quad \text{(A.3.17)}$$

$|\delta_1 - \delta_2 + s_1 - s_2|$ は (4.3), (4.7), (A.3.17) を順に用いて,

$$\begin{aligned}&|\delta_1 - \delta_2 + s_1 - s_2| \\ &\leq |\delta_1| + |\delta_2| + |s_1| + |s_2| \\ &\leq (c_1\mathbf{u} + d_1\mathbf{u}^2)f_1 + (c_2\mathbf{u} + d_2\mathbf{u}^2)f_2 + e_1\mathbf{u}_S + e_2\mathbf{u}_S \\ &\leq \alpha(f_1 + f_2) + (e_1 + e_2)\mathbf{u}_S \\ &\leq (1+\mathbf{u})\alpha\mathrm{fl}(f_1 + f_2) + (e_1 + e_2)\mathbf{u}_S\end{aligned}$$

$$= (1+\mathbf{u})^{-2}(1+\mathbf{u})^{3}\alpha\mathrm{fl}(f_1+f_2) + (e_1+e_2)\mathbf{u}_S =: U$$

と上限を取れる. ここで, (4.7) より $(1+\mathbf{u})^3\alpha \leq \varphi \in \mathbb{F}$ であり, (2.3) を用いて U の上限は以下のように求める.

$$\begin{aligned}
U &\leq (1+\mathbf{u})^{-2}\varphi\ \mathrm{fl}(f_1+f_2) + (e_1+e_2)\mathbf{u}_S \\
&\leq (1+\mathbf{u})^{-2}\left((1+\mathbf{u})\mathrm{fl}\left(\varphi\cdot(f_1+f_2)\right) + \frac{\mathbf{u}_S}{2}\right) + (e_1+e_2)\mathbf{u}_S \\
&= (1+\mathbf{u})^{-1}\mathrm{fl}\left(\varphi\cdot(f_1+f_2)\right) + \left(e_1+e_2+\frac{1}{2(1+\mathbf{u})^2}\right)\mathbf{u}_S \\
&< (1+\mathbf{u})^{-1}\mathrm{fl}\left(\varphi\cdot(f_1+f_2)\right) + (e_1+e_2+1)\mathbf{u}_S - \frac{\mathbf{u}_S}{2} \qquad \text{(A.3.18)}
\end{aligned}$$

ここで, (A.3.18) における $\mathbf{u}_S$ の項の上限を求める.

$$\begin{aligned}
&(e_1+e_2+1)\mathbf{u}_S - \frac{\mathbf{u}_S}{2} \\
&= (1+\mathbf{u})^{-4}(1+\mathbf{u})^{4}(e_1+e_2+1)\mathbf{u}_S - \frac{\mathbf{u}_S}{2} \\
&< (1+\mathbf{u})^{-3}(1+6\mathbf{u})(\mathrm{fl}(e_1+e_2)+1)\mathbf{u}_S - \frac{\mathbf{u}_S}{2} \\
&\leq (1+\mathbf{u})^{-2}(1+6\mathbf{u})\mathrm{fl}\big((e_1+e_2)+1\big)\mathbf{u}_S - \frac{\mathbf{u}_S}{2} \\
&\leq (1+\mathbf{u})^{-1}\mathrm{fl}\left((1+6\mathbf{u})\cdot((e_1+e_2)+1)\right)\mathbf{u}_S - \frac{\mathbf{u}_S}{2} \\
&\leq (1+\mathbf{u})^{-1}\left(\mathrm{fl}\left((1+6\mathbf{u})\cdot((e_1+e_2)+1)\mathbf{u}_S\right) + \frac{\mathbf{u}_S}{2}\right) - \frac{\mathbf{u}_S}{2} \\
&= (1+\mathbf{u})^{-1}\mathrm{fl}\left((1+6\mathbf{u})\cdot((e_1+e_2)+1)\mathbf{u}_S\right) + \frac{1}{2(1+\mathbf{u})}\mathbf{u}_S - \frac{\mathbf{u}_S}{2} \\
&< (1+\mathbf{u})^{-1}\mathrm{fl}\left((1+6\mathbf{u})\cdot((e_1+e_2)+1)\mathbf{u}_S\right) \qquad \text{(A.3.19)}
\end{aligned}$$

上式の導出について, 1 行目から 3 行目は (2.4) を用いた. 2 行目では $\mathbb{F} \not\ni (1+\mathbf{u})^4 < 1+6\mathbf{u} \in \mathbb{F}$ の関係から上限を取った. 3 行目から 4 行目は (2.3) を用いた. ここでは 1 以上の数どうしの積を計算したため, アンダーフローが発生しない. そこで $\mathbf{u}_S$ の項がないものを利用した. 4 行目から 5 行目は (2.5) を用いた. 最後に (A.3.18) の上限を (A.3.19) を用いて求め, (2.4) を用いて和の上限を取ることで次の結果を得る.

$$\begin{aligned}
&|\delta_1 - \delta_2 + s_1 - s_2| \\
&< (1+\mathbf{u})^{-1}\mathrm{fl}\left(\varphi\cdot(f_1+f_2)\right) + (1+\mathbf{u})^{-1}\mathrm{fl}\left((1+6\mathbf{u})\cdot((e_1+e_2)+1)\mathbf{u}_S\right) \\
&= (1+\mathbf{u})^{-1}(\mathrm{fl}\left(\varphi\cdot(f_1+f_2)\right) + \mathrm{fl}\left((1+6\mathbf{u})\cdot((e_1+e_2)+1)\mathbf{u}_S\right)) \\
&\leq \mathrm{fl}\left(\varphi\cdot(f_1+f_2) + (1+6\mathbf{u})\cdot((e_1+e_2)+1)\mathbf{u}_S\right)
\end{aligned}$$

□

補題 4.3 の証明

$c\mathbf{u} = \mathrm{fl}(c\mathbf{u})$, $4c = \mathrm{fl}(4c)$ であり, $\mathbf{u}$ の 3 次以上の係数に現れる c, d を $c < \mathbf{u}^{-1}$, $d < \mathbf{u}^{-1}$ と評価すると, 以下のようになる.

$$
\begin{aligned}
&(1+\mathbf{u})^3(c\mathbf{u}+d\mathbf{u}^2)\\
&= \frac{1+\mathbf{u}}{1+\mathbf{u}}c\mathbf{u} + (3c+d)\mathbf{u}^2 + (3c+3d)\mathbf{u}^3 + (c+3d)\mathbf{u}^4 + d\mathbf{u}^5\\
&= \frac{c\mathbf{u}}{1+\mathbf{u}} + \frac{c\mathbf{u}^2}{1+\mathbf{u}} + (3c+d)\mathbf{u}^2 + (3c+3d)\mathbf{u}^3 + (c+3d)\mathbf{u}^4 + d\mathbf{u}^5\\
&\le \frac{c\mathbf{u}}{1+\mathbf{u}} + (4c+d)\mathbf{u}^2 + (3c+3d)\mathbf{u}^3 + (c+3d)\mathbf{u}^4 + d\mathbf{u}^5\\
&= \frac{c\mathbf{u}}{1+\mathbf{u}} + \frac{(1+\mathbf{u})^3}{(1+\mathbf{u})^3}\left((4c+d)\mathbf{u}^2 + (3c+3d)\mathbf{u}^3 + (c+3d)\mathbf{u}^4 + d\mathbf{u}^5\right)\\
&< \frac{c\mathbf{u}}{1+\mathbf{u}} + \frac{1+4\mathbf{u}}{(1+\mathbf{u})^3}\left((4c+d+6)\mathbf{u}^2 + 4\mathbf{u}^3 + \mathbf{u}^4\right)\\
&= \frac{c\mathbf{u}}{1+\mathbf{u}} + \frac{1}{(1+\mathbf{u})^3}\left((4c+d+6)\mathbf{u}^2 + (16c+4d+28)\mathbf{u}^3 + 17\mathbf{u}^4 + 4\mathbf{u}^5\right)\\
&< \frac{c\mathbf{u}}{1+\mathbf{u}} + \frac{(4c+d+27)\mathbf{u}^2}{(1+\mathbf{u})^3}\\
&= \frac{\mathrm{fl}(c\mathbf{u})}{1+\mathbf{u}} + \frac{(\mathrm{fl}(4c)+d+27)\mathbf{u}^2}{(1+\mathbf{u})^3}\\
&\le \frac{\mathrm{fl}(c\mathbf{u})}{1+\mathbf{u}} + \frac{(\mathrm{fl}(4c+d)+27)\,\mathbf{u}^2}{(1+\mathbf{u})^2}\\
&\le \frac{\mathrm{fl}(c\mathbf{u})}{1+\mathbf{u}} + \frac{\mathrm{fl}\left(((4c+d)+27)\,\mathbf{u}^2\right)}{1+\mathbf{u}}\\
&\le \mathrm{fl}\left(c\mathbf{u} + ((4c+d)+27)\,\mathbf{u}^2\right)
\end{aligned}
$$

5 行目から 6 行目は $(1+\mathbf{u})^3 < 1+4\mathbf{u}$ の関係, 最後の 3 つの式変形は (2.4) を利用した. □

定理 4.4 の証明

仮定 (4.9) と補題補題 2.7 より, 以下の不等式を得る.

$$
\begin{aligned}
&\mathrm{fl}(|x_1 - x_2|)\\
\ge&\mathrm{fl}\left(\varphi\cdot(f_1+f_2) + (1+6\mathbf{u})\cdot((e_1+e_2)+1)\mathbf{u}_S\right) + \mathbf{u}\cdot\mathrm{ufp}(\mathrm{fl}(x_1-x_2)) \quad \text{(A.3.20)}
\end{aligned}
$$

$\mathrm{fl}(x_1-x_2)$ の演算における丸め誤差 δ の絶対値の最大値は, 補題 2.4 より $\mathbf{u}\cdot\mathrm{ufp}(\mathrm{fl}(x_1-x_2))$ である. (A.3.20) の右辺の下限を (4.8) より求める.

$$
\begin{aligned}
&\mathrm{fl}(|x_1 - x_2|)\\
&\ge \mathrm{fl}(\varphi\cdot(f_1+f_2) + (1+6\mathbf{u})\cdot((e_1+e_2)+1)\mathbf{u}_S) + \mathbf{u}\cdot\mathrm{ufp}(\mathrm{fl}(x_1-x_2))\\
&> |\delta_1 - \delta_2 + s_1 - s_2| + |\delta|\\
&\ge |\delta_1 - \delta_2 + s_1 - s_2 + \delta|
\end{aligned}
$$

よって (4.6) を満たすため, $\mathrm{fl}(x_1-x_2)$ と t_1-t_2 の符号は同じとなる. □

定理 4.5 の証明

`realmax` を $\mathbb{F}$ に属する最大数とする. まず,

- NaN を含んだ論理式 $p_1 > p_2$ の評価はすべて偽となる
- p_2 が Inf の場合, すべての p_1 に対して論理式 $p_1 > p_2$ は偽となる
- $p_1, p_2 \in \mathbb{F}$ のとき, 定理 4.4 より結果は保証される
- p_1, p_2 は $-\mathrm{Inf}$ となることはない
- 仮定より x_i が $\pm\mathrm{Inf}, \mathrm{NaN}$ のいずれかならば, f_i も $\pm\mathrm{Inf}, \mathrm{NaN}$ になる. このとき $p_2 \in \{\mathrm{Inf}, \mathrm{NaN}\}$ となることから論理式は偽となる

ため, $p_1 = \mathrm{Inf}$, また $p_2, x_1, x_2 \in \mathbb{F}$ のときのみを考えればよい. $p_1 = \mathrm{Inf}$ より $|x_1 - x_2| >$ `realmax` であり, $p_2 \in \mathbb{F}$, (4.8) を用いれば

$$|x_1 - x_2| > \texttt{realmax} \geq p_2 > |\delta_1 - \delta_2 + s_1 - s_2|$$

を満たすため, (4.5) より $x_1 - x_2$ の符号と $t_1 - t_2$ の符号は等しい. また $x_1, x_2 \in \mathbb{F}$ であるため, $x_1 - x_2$ と $\mathrm{fl}(x_1 - x_2)$ の符号は等しい. □

補題 4.6 の証明

$\frac{1}{2}\mathbf{u}^{-1} < (1 + 6\mathbf{u})(\frac{1}{2}\mathbf{u}^{-1} - \frac{5}{2}) = \frac{1}{2}\mathbf{u}^{-1}(1 + \mathbf{u} - 30\mathbf{u}^2) < \frac{1}{2}\mathbf{u}^{-1}(1 + \mathbf{u})$ より $\mathrm{fl}\left((1+6\mathbf{u}) \cdot (\frac{1}{2}\mathbf{u}^{-1} - \frac{5}{2})\right) = \frac{1}{2}\mathbf{u}^{-1}$ となる. これより, 次のように上限を求めることができる.

$$\begin{aligned}
&\mathrm{fl}\left((1+6\mathbf{u}) \cdot ((e_1+e_2)+1)\mathbf{u}_S\right) \\
&\leq \mathrm{fl}\left((1+6\mathbf{u}) \cdot \left(\left(\frac{1}{2}\mathbf{u}^{-1} - \frac{7}{2}\right) + 1\right)\mathbf{u}_S\right) \\
&= \mathrm{fl}\left((1+6\mathbf{u}) \cdot \left(\frac{1}{2}\mathbf{u}^{-1} - \frac{5}{2}\right)\mathbf{u}_S\right) \\
&= \mathrm{fl}\left(\frac{1}{2}\mathbf{u}^{-1}\mathbf{u}_S\right) = \mathbf{u}_N
\end{aligned}$$

最後の式変形は (2.1) を利用した. よって, 補題は成り立つ. □

A.4 5 章の定理の証明

定理 5.2 の証明

補題 2.11 より, $|x - a| \leq X$ が成り立つ. もし $a = 0$ であれば, $|x| \leq X$, $Y \leq 0$ であるから, $\mathrm{fl}(k(Y - r)) < 0$ となり, (5.3) が満たされることはない. よって, $a \neq 0$ として証明を続ける. ここで,

$$\frac{|x-a|}{|a|} < k \Leftrightarrow |x - a| < k|a|.$$

と式変形できる. 定理 2.2 を利用し, $|a|$ の下限を次のように評価する.

$$|a| \geq |x| - |x-a| \geq |x| - X \geq Y - r = \mathrm{fl}(Y-r), \quad Y - r \in \mathbb{F}.$$

よって, $X < k\mathrm{fl}(Y-r)$ であれば $|x-a| < k|a|$ となる. 補題 2.8 より, (5.3) が満たされるならば, $X < k\mathrm{fl}(Y-r)$ が成り立つ. よって, 命題は成り立つ. □

定理 5.3 の証明

(2.10) より以下を得る.

$$\begin{aligned} |x| &> \mathrm{fl}((c_1\mathbf{u} + (d_1 + 3\mathrm{ufp}(c_1))\mathbf{u}^2)\cdot f_1) + (e_1+1)\mathbf{u}_S \\ &> \mathrm{fl}((c_1\mathbf{u} + (d_1 + 3\mathrm{ufp}(c_1))\mathbf{u}^2)\cdot f_1) + \frac{\mathbf{u}_S}{2} + e_1\mathbf{u}_S \\ &> (c_1\mathbf{u} + d_1\mathbf{u}^2)f_1 + e_1\mathbf{u}_S \end{aligned}$$

最後の式変形については, 補題 2.11 を利用した. □

定理 5.5 の証明

$a = 0$ のときは $Z < 0$ であり, 定理 5.2 と同様にして, (5.4) は満たされない. $a \neq 0$ の場合を考える. ここで

$$\frac{|x-a|}{|a|} < k \Leftrightarrow |x-a| < k|a|.$$

と式変形する. 補題 2.11 より, $|x-a| \leq Y$ である. また, 定理 2.2 より $|a|$ の下限を以下のように求めることができる.

$$|a| \geq |x| - |x-a| \geq |x| - Y \geq Z - r = \mathrm{fl}(Z-r)$$

補題 2.8 より, (5.4) が満たされれば, $X < k\mathrm{fl}(Z-s)$ が成り立つ. すなわち, $|x-a| < k|a|$ が成り立つ. □

定理 5.7 の証明

定理 5.2, 定理 5.5 と同様にして, $a \neq 0$ に対して $|x-a| < k|a|$ を考える. 補題 2.11 より $|x-a| \leq X$ であり, 定理 2.2 より $|t|$ の下限は次のようになる.

$$|a| \geq |x| - |x-a| \geq |x| - X \geq Y - r = \mathrm{fl}(Y-r)$$

補題 2.8 より, (5.5) が満たされれば, $X < k\mathrm{fl}(Y-r)$ が成り立つ. すなわち, $|x-a| < k|a|$ が成り立つ. □

定理 5.8 の証明

$\mathrm{fl}(x \pm y)$, $a \pm b$ に対する誤差上限を直接評価する. ただし, $j := \arg\max_i \left(c_i\mathbf{u} + d_i\mathbf{u}^2\right)$, $c := c_j$, $d := d_j$, $\alpha = (c_1+c_2)\mathbf{u} + (d_1+d_2)\mathbf{u}^2$ とし, $\pm$ は複号同順とする. また, float で評価する際は補題 2.6 を利用する.

$|\mathrm{fl}(x \pm y) - (a \pm b)|$

$$
\begin{aligned}
=&|\mathrm{fl}(x \pm y) - (x \pm y) + (x - a) \pm (y - b)| \\
\leq&|\mathrm{fl}(x \pm y) - (x \pm y)| + |x - a| + |y - b| \\
\leq&\mathbf{u}\mathrm{fl}(|x \pm y|) + \left(\left(c_1\mathbf{u} + d_1\mathbf{u}^2\right) f_1 + e_1\mathbf{u}_S\right) + \left((c_2\mathbf{u} + d_2\mathbf{u}^2)f_2 + e_2\mathbf{u}_S\right) \\
\leq&\mathbf{u}\mathrm{fl}(|x| + |y|) + \left(\left(c_1\mathbf{u} + d_1\mathbf{u}^2\right) f_1 + e_1\mathbf{u}_S\right) + \left((c_2\mathbf{u} + d_2\mathbf{u}^2)f_2 + e_2\mathbf{u}_S\right) \\
\leq&\mathbf{u}\mathrm{fl}(f_1 + f_2) + \left(\left(c_1\mathbf{u} + d_1\mathbf{u}^2\right) f_1 + (c_2\mathbf{u} + d_2\mathbf{u}^2)f_2\right) + (e_1 + e_2)\mathbf{u}_S \\
\leq&\mathbf{u}\mathrm{fl}(f_1 + f_2) + \alpha(f_1 + f_2) + (e_1 + e_2)\mathbf{u}_S \\
\leq&\mathbf{u}\mathrm{fl}(f_1 + f_2) + \alpha(1 + \mathbf{u})\mathrm{fl}(f_1 + f_2) + (e_1 + e_2)\mathbf{u}_S \\
=&\left(\mathbf{u} + \alpha(1 + \mathbf{u})\right)\mathrm{fl}(f_1 + f_2) + (e_1 + e_2)\mathbf{u}_S \\
=&\left((c + 1)\mathbf{u} + (c + d + 1)\mathbf{u}^2 + d\mathbf{u}^3\right)\mathrm{fl}(f_1 + f_2) + (e_1 + e_2)\mathbf{u}_S \\
=&\left(\frac{1 + \mathbf{u}}{1 + \mathbf{u}}(c + 1)\mathbf{u} + (c + d + 1)\,\mathbf{u}^2 + d\mathbf{u}^3\right)\mathrm{fl}(f_1 + f_2) + (e_1 + e_2)\mathbf{u}_S \\
=&\left(\frac{1}{1 + \mathbf{u}}(c + 1)\mathbf{u} + \left(c + d + 1 + \frac{c + 1}{1 + \mathbf{u}}\right)\mathbf{u}^2 + d\mathbf{u}^3\right)\mathrm{fl}(f_1 + f_2) + (e_1 + e_2)\mathbf{u}_S \\
<&\left(\frac{1}{1 + \mathbf{u}}(c + 1)\mathbf{u} + (2c + d + 2)\,\mathbf{u}^2 + d\mathbf{u}^3\right)\mathrm{fl}(f_1 + f_2) + (e_1 + e_2)\mathbf{u}_S \\
\leq&\left(\mathrm{fl}(c + 1)\mathbf{u} + \left((2c + d + 2)\,\mathbf{u}^2 + d\mathbf{u}^3\right)\right)\mathrm{fl}(f_1 + f_2) + (e_1 + e_2)\mathbf{u}_S \\
=&\left(\mathrm{fl}(c + 1)\mathbf{u} + \frac{(1 + \mathbf{u})^2}{(1 + \mathbf{u})^2}\left((2c + d + 2)\,\mathbf{u}^2 + d\mathbf{u}^3\right)\right)\mathrm{fl}(f_1 + f_2) + (e_1 + e_2)\mathbf{u}_S \\
=&\left(\mathrm{fl}(c + 1)\mathbf{u} + \frac{1}{(1 + \mathbf{u})^2}\left((2c + d + 2)\,\mathbf{u}^2 + (3c + 3d + 3)\,\mathbf{u}^3 + (c + 3d + 1)\,\mathbf{u}^4 + d\mathbf{u}^5\right)\right) \\
&\mathrm{fl}(f_1 + f_2) + (e_1 + e_2)\mathbf{u}_S \\
\leq&\left(\mathrm{fl}(c + 1)\mathbf{u}\right. \\
&\left.+\frac{1}{(1 + \mathbf{u})^2}\left((2c + d + 2)\,\mathbf{u}^2 + (3U + 3U + 3)\,\mathbf{u}^3 + (U + 3U + 1)\,\mathbf{u}^4 + U\mathbf{u}^5\right)\right) \\
&\mathrm{fl}(f_1 + f_2) + (e_1 + e_2)\mathbf{u}_S \\
=&\left(\mathrm{fl}(c + 1)\mathbf{u} + \frac{1}{(1 + \mathbf{u})^2}\left((2c + d + 8)\,\mathbf{u}^2 + \mathbf{u}^3 - 2\mathbf{u}^4 - \mathbf{u}^5\right)\right)\mathrm{fl}(f_1 + f_2) + (e_1 + e_2)\mathbf{u}_S \\
<&\left(\mathrm{fl}(c + 1)\mathbf{u} + \frac{1}{(1 + \mathbf{u})^2}\left(2c + d + 9\right)\mathbf{u}^2\right)\mathrm{fl}(f_1 + f_2) + (e_1 + e_2)\mathbf{u}_S \\
=&\left(\mathrm{fl}(c + 1)\mathbf{u} + \frac{1}{(1 + \mathbf{u})^2}\left(\mathrm{fl}(2c) + d + 9\right)\mathbf{u}^2\right)\mathrm{fl}(f_1 + f_2) + (e_1 + e_2)\mathbf{u}_S \\
\leq&\left(\mathrm{fl}(c + 1)\mathbf{u} + \frac{1}{(1 + \mathbf{u})^2}(1 + \mathbf{u})^2\mathrm{float}\left(2c + d + 9\right)\mathbf{u}^2\right)\mathrm{fl}(f_1 + f_2) + (e_1 + e_2)\mathbf{u}_S \\
=&\left(\mathrm{fl}(c + 1)\mathbf{u} + \mathrm{float}\left(2c + d + 9\right)\mathbf{u}^2\right)\mathrm{fl}(f_1 + f_2) + (e_1 + e_2)\mathbf{u}_S \\
=&\left(\mathrm{fl}(c + 1)\mathbf{u} + \mathrm{float}\left(2c + d + 9\right)\mathbf{u}^2\right)\mathrm{fl}(f_1 + f_2) + \frac{(1 + \mathbf{u})^2}{(1 + \mathbf{u})^2}(e_1 + e_2)\mathbf{u}_S \\
<&\left(\mathrm{fl}(c + 1)\mathbf{u} + \mathrm{float}\left(2c + d + 9\right)\mathbf{u}^2\right)\mathrm{fl}(f_1 + f_2) + \frac{1}{(1 + \mathbf{u})^2}(1 + 2\mathbf{u})(e_1 + e_2)\mathbf{u}_S
\end{aligned}
$$

$$\leq \left(\mathrm{fl}(c+1)\mathbf{u} + \mathrm{float}\left(2c+d+9\right)\mathbf{u}^2\right)\mathrm{fl}(f_1+f_2) + \frac{1}{(1+\mathbf{u})^2}\mathrm{fl}(1+2\mathbf{u})(1+\mathbf{u})\mathrm{fl}(e_1+e_2)\mathbf{u}_S$$

$$= \left(\mathrm{fl}(c+1)\mathbf{u} + \mathrm{float}\left(2c+d+9\right)\mathbf{u}^2\right)\mathrm{fl}(f_1+f_2) + \frac{1}{1+\mathbf{u}}\mathrm{fl}(1+2\mathbf{u})\mathrm{fl}(e_1+e_2)\mathbf{u}_S$$
$$\leq \left(\mathrm{fl}(c+1)\mathbf{u} + \mathrm{float}\left(2c+d+9\right)\mathbf{u}^2\right)\mathrm{fl}(f_1+f_2) + \frac{1}{1+\mathbf{u}}(1+\mathbf{u})\mathrm{fl}\left((1+2\mathbf{u})\cdot(e_1+e_2)\right)\mathbf{u}_S$$

$$= \left(\mathrm{fl}(c+1)\mathbf{u} + \mathrm{float}\left(2c+d+9\right)\mathbf{u}^2\right)\mathrm{fl}(f_1+f_2) + \mathrm{fl}\left((1+2\mathbf{u})\cdot(e_1+e_2)\right)\mathbf{u}_S$$
$$\leq \left(\mathrm{fl}(c+1)\mathbf{u} + \mathrm{float}\left(2c+d+9\right)\mathbf{u}^2\right)\mathrm{fl}(f_1+f_2) + \lceil\mathrm{fl}\left((1+2\mathbf{u})\cdot(e_1+e_2)\right)\rceil\mathbf{u}_S$$

□

定理 5.9 の証明

誤差解析の見やすさのため, 次の記号を利用する.

$$\begin{aligned}
&U = \mathbf{u}^{-1} - 1 \geq c, d,\\
&X = (c_1+c_2)\,\mathbf{u} + (c_1c_2+d_1+d_2)\,\mathbf{u}^2 + (c_1d_2+c_2d_1)\,\mathbf{u}^3 + d_1d_2\mathbf{u}^4,\\
&Y = \left(1+c_1\mathbf{u}+d_1\mathbf{u}^2\right)f_1e_2 + \left(1+c_2\mathbf{u}+d_2\mathbf{u}^2\right)f_2e_1 + \frac{1}{2} + e_1e_2\mathbf{u}_S.
\end{aligned}$$

これを利用して, $\mathrm{fl}(x\cdot y)$, ab に対する誤差上限を評価する. また, float で評価する際は補題 2.6 を利用する.

$$\begin{aligned}
&|\mathrm{fl}(x\cdot y) - (ab)|\\
=&|\mathrm{fl}(x\cdot y) - xy + xy - ab|\\
\leq&|\mathrm{fl}(x\cdot y) - xy| + |xy - ab|\\
=&|\mathrm{fl}(x\cdot y) - xy| + |xy - (x+\delta_1)(y+\delta_2)|\\
=&|\mathrm{fl}(x\cdot y) - xy| + |-x\delta_2 - \delta_1 y - \delta_1\delta_2|\\
\leq&\mathbf{u}\mathrm{fl}(|x\cdot y|) + \frac{\mathbf{u}_S}{2} + |x||\delta_2| + |\delta_1||y| + |\delta_1||\delta_2|\\
\leq&\mathbf{u}\mathrm{fl}(f_1\cdot f_2) + \frac{\mathbf{u}_S}{2} + f_1\left(\left(c_2\mathbf{u}+d_2\mathbf{u}^2\right)f_2 + e_2\mathbf{u}_S\right) + \left(\left(c_1\mathbf{u}+d_1\mathbf{u}^2\right)f_1 + e_1\mathbf{u}_S\right)f_2\\
&+ \left(\left(c_1\mathbf{u}+d_1\mathbf{u}^2\right)f_1 + e_1\mathbf{u}_S\right)\left(\left(c_2\mathbf{u}+d_2\mathbf{u}^2\right)f_2 + e_2\mathbf{u}_S\right)\\
=&\mathbf{u}\mathrm{fl}(f_1\cdot f_2) + \frac{\mathbf{u}_S}{2} + \left(c_2\mathbf{u}+d_2\mathbf{u}^2\right)f_1f_2 + f_1e_2\mathbf{u}_S + \left(c_1\mathbf{u}+d_1\mathbf{u}^2\right)f_1f_2 + f_2e_1\mathbf{u}_S\\
&+ \left(c_1\mathbf{u}+d_1\mathbf{u}^2\right)\left(c_2\mathbf{u}+d_2\mathbf{u}^2\right)f_1f_2 + \left(c_1\mathbf{u}+d_1\mathbf{u}^2\right)f_1e_2\mathbf{u}_S + \left(c_2\mathbf{u}+d_2\mathbf{u}^2\right)e_1f_2\mathbf{u}_S\\
&+ e_1e_2\mathbf{u}_S^2\\
=&\mathbf{u}\mathrm{fl}(f_1\cdot f_2) + \left((c_1+c_2)\,\mathbf{u} + (c_1c_2+d_1+d_2)\,\mathbf{u}^2 + (c_1d_2+c_2d_1)\,\mathbf{u}^3 + d_1d_2\mathbf{u}^4\right)f_1f_2\\
&+ \left(\left(1+c_1\mathbf{u}+d_1\mathbf{u}^2\right)f_1e_2 + \left(1+c_2\mathbf{u}+d_2\mathbf{u}^2\right)f_2e_1 + \frac{1}{2} + e_1e_2\mathbf{u}_S\right)\mathbf{u}_S\\
=&\mathbf{u}\mathrm{fl}(f_1\cdot f_2) + Xf_1f_2 + Y\mathbf{u}_S\\
\leq&\mathbf{u}\mathrm{fl}(f_1\cdot f_2) + X\left((1+\mathbf{u})\mathrm{fl}(f_1\cdot f_2) + \frac{\mathbf{u}_S}{2}\right) + Y\mathbf{u}_S\\
=&\left(X(1+\mathbf{u})+\mathbf{u}\right)\mathrm{fl}(f_1\cdot f_2) + (Y + X/2)\,\mathbf{u}_S
\end{aligned}$$

$$
\begin{aligned}
&= \big((c_1+c_2+1)\,\mathbf{u} + (c_1+c_2+d_1+d_2+c_1c_2)\,\mathbf{u}^2 + (d_1+d_2+c_1c_2+c_1d_2+c_2d_1)\,\mathbf{u}^3 \\
&\quad + (c_1d_2+c_2d_1+d_1d_2)\,\mathbf{u}^4 + d_1d_2\mathbf{u}^5\big)\,\mathrm{fl}\,(f_1\cdot f_2) + (Y+X/2)\,\mathbf{u}_S \\
&= \Bigg(\frac{(1+\mathbf{u})^2}{(1+\mathbf{u})^2}\,(c_1+c_2+1)\,\mathbf{u} + (c_1+c_2+d_1+d_2+c_1c_2)\,\mathbf{u}^2 \\
&\quad + (d_1+d_2+c_1c_2+c_1d_2+c_2d_1)\,\mathbf{u}^3 + (c_1d_2+c_2d_1+d_1d_2)\,\mathbf{u}^4 + d_1d_2\mathbf{u}^5\Bigg)\,\mathrm{fl}\,(f_1\cdot f_2) \\
&\quad + (Y+X/2)\,\mathbf{u}_S \\
&= \Bigg(\frac{1}{(1+\mathbf{u})^2}\,(c_1+c_2+1)\,\mathbf{u} + \left(c_1+c_2+d_1+d_2+c_1c_2+2\frac{c_1+c_2+1}{(1+\mathbf{u})^2}\right)\mathbf{u}^2 \\
&\quad + \left(d_1+d_2+c_1c_2+c_1d_2+c_2d_1+\frac{c_1+c_2+1}{(1+\mathbf{u})^2}\right)\mathbf{u}^3 \\
&\quad + (c_1d_2+c_2d_1+d_1d_2)\,\mathbf{u}^4 + d_1d_2\mathbf{u}^5\Bigg)\,\mathrm{fl}\,(f_1\cdot f_2) \\
&\quad + (Y+X/2)\,\mathbf{u}_S \\
&\le \Bigg(\frac{1}{(1+\mathbf{u})^2}(1+\mathbf{u})^2\mathrm{float}\,(c_1+c_2+1)\,\mathbf{u} + \left(c_1+c_2+d_1+d_2+c_1c_2+2\frac{c_1+c_2+1}{(1+\mathbf{u})^2}\right)\mathbf{u}^2 \\
&\quad + \left(d_1+d_2+c_1c_2+c_1d_2+c_2d_1+\frac{c_1+c_2+1}{(1+\mathbf{u})^2}\right)\mathbf{u}^3 \\
&\quad + (c_1d_2+c_2d_1+d_1d_2)\,\mathbf{u}^4 + d_1d_2\mathbf{u}^5\Bigg)\,\mathrm{fl}\,(f_1\cdot f_2) \\
&\quad + (Y+X/2)\,\mathbf{u}_S \\
&= \Bigg(\mathrm{float}\,(c_1+c_2+1)\,\mathbf{u} + \left(c_1+c_2+d_1+d_2+c_1c_2+2\frac{c_1+c_2+1}{(1+\mathbf{u})^2}\right)\mathbf{u}^2 \\
&\quad + \left(d_1+d_2+c_1c_2+c_1d_2+c_2d_1+\frac{c_1+c_2+1}{(1+\mathbf{u})^2}\right)\mathbf{u}^3 \\
&\quad + (c_1d_2+c_2d_1+d_1d_2)\,\mathbf{u}^4 + d_1d_2\mathbf{u}^5\Bigg)\,\mathrm{fl}\,(f_1\cdot f_2) \\
&\quad + (Y+X/2)\,\mathbf{u}_S \\
&= \Bigg(\mathrm{float}\,(c_1+c_2+1)\,\mathbf{u} + \frac{(1+\mathbf{u})^6}{(1+\mathbf{u})^6}\left(\left(c_1+c_2+d_1+d_2+c_1c_2+2\frac{c_1+c_2+1}{(1+\mathbf{u})^2}\right)\mathbf{u}^2\right. \\
&\quad + \left(d_1+d_2+c_1c_2+c_1d_2+c_2d_1+\frac{c_1+c_2+1}{(1+\mathbf{u})^2}\right)\mathbf{u}^3 \\
&\quad + (c_1d_2+c_2d_1+d_1d_2)\,\mathbf{u}^4 + d_1d_2\mathbf{u}^5\Bigg)\,\mathrm{fl}\,(f_1\cdot f_2) \\
&\quad + (Y+X/2)\,\mathbf{u}_S
\end{aligned}
$$

$$
\begin{aligned}
=&\Big(\text{float}\,(c_1+c_2+1)\,\mathbf{u}+\frac{1}{(1+\mathbf{u})^6}\left((3c_1+3c_2+d_1+d_2+c_1c_2+2)\,\mathbf{u}^2\right.\\
&+(15c_1+15c_2+7d_1+7d_2+7c_1c_2+c_1d_2+c_2d_1+9)\,\mathbf{u}^3\\
&+(31c_1+31c_2+21d_1+21d_2+21c_1c_2+7c_1d_2+7c_2d_1+d_1d_2+16)\,\mathbf{u}^4\\
&+(34c_1+34c_2+35d_1+35d_2+35c_1c_2+21c_1d_2+21c_2d_1+7d_1d_2+14)\,\mathbf{u}^5\\
&+(21c_1+21c_2+35d_1+35d_2+35c_1c_2+35c_1d_2+35c_2d_1+21d_1d_2+6)\,\mathbf{u}^6\\
&+(7c_1+7c_2+21d_1+21d_2+21c_1c_2+35c_1d_2+35c_2d_1+35d_1d_2+1)\,\mathbf{u}^7\\
&+(c_1+c_2+7d_1+7d_2+7c_1c_2+21c_1d_2+21c_2d_1+35d_1d_2)\,\mathbf{u}^8\\
&+(d_1+d_2+c_1c_2+7c_1d_2+7c_2d_1+21d_1d_2)\,\mathbf{u}^9\\
&\left.+(c_1d_2+c_2d_1+7d_1d_2)\,\mathbf{u}^{10}+d_1d_2\mathbf{u}^{11}\right)\Big)\,\text{fl}\,(f_1\cdot f_2)\\
&+(Y+X/2)\,\mathbf{u}_S\\
\le&\Big(\text{float}\,(c_1+c_2+1)\,\mathbf{u}+\frac{1}{(1+\mathbf{u})^6}\left((3c_1+3c_2+d_1+d_2+c_1c_2+2)\,\mathbf{u}^2\right.\\
&+(15c_1+15c_2+7d_1+7d_2+7c_1c_2+c_1d_2+c_2d_1+9)\,\mathbf{u}^3\\
&+\left(31U+31U+21U+21U+21U^2+7U^2+7U^2+U^2+16\right)\mathbf{u}^4\\
&+\left(34U+34U+35U+35U+35U^2+21U^2+21U^2+7U^2+14\right)\mathbf{u}^5\\
&+\left(21U+21U+35U+35U+35U^2+35U^2+35U^2+21U^2+6\right)\mathbf{u}^6\\
&+\left(7U+7U+21U+21U+21U^2+35U^2+35U^2+35U^2+1\right)\mathbf{u}^7\\
&+\left(U+U+7U+7U+7U^2+21U^2+21U^2+35U^2\right)\mathbf{u}^8\\
&+\left(U+U+U^2+7U^2+7U^2+21U^2\right)\mathbf{u}^9\\
&\left.+\left(U^2+U^2+7U^2\right)\mathbf{u}^{10}+U^2\mathbf{u}^{11}\right)\Big)\,\text{fl}\,(f_1\cdot f_2)\\
&+(Y+X/2)\,\mathbf{u}_S\\
=&\Big(\text{float}\,(c_1+c_2+1)\,\mathbf{u}+\frac{1}{(1+\mathbf{u})^6}\left((3c_1+3c_2+d_1+d_2+c_1c_2+38)\,\mathbf{u}^2\right.\\
&+(15c_1+15c_2+7d_1+7d_2+7c_1c_2+c_1d_2+c_2d_1+125)\,\mathbf{u}^3\\
&\left.+44\mathbf{u}^4-54\mathbf{u}^5-92\mathbf{u}^6-45\mathbf{u}^7+7\mathbf{u}^8+17\mathbf{u}^9+7\mathbf{u}^{10}+\mathbf{u}^{11}\right)\Big)\,\text{fl}\,(f_1\cdot f_2)\\
&+(Y+X/2)\,\mathbf{u}_S\\
\le&\Big(\text{float}\,(c_1+c_2+1)\,\mathbf{u}+\frac{1}{(1+\mathbf{u})^6}\left((3c_1+3c_2+d_1+d_2+c_1c_2+38)\,\mathbf{u}^2\right.\\
&+(5c_1+15c_2+7d_1+7d_2+7(c_1U+Uc_2)+Ud_2+Ud_1+125)\,\mathbf{u}^3\\
&\left.+44\mathbf{u}^4-54\mathbf{u}^5-92\mathbf{u}^6-45\mathbf{u}^7+7\mathbf{u}^8+17\mathbf{u}^9+7\mathbf{u}^{10}+\mathbf{u}^{11}\right)\Big)\,\text{fl}\,(f_1\cdot f_2)\\
&+(Y+X/2)\,\mathbf{u}_S\\
=&\Big(\text{float}\,(c_1+c_2+1)\,\mathbf{u}+\frac{1}{(1+\mathbf{u})^6}\left((10c_1+10c_2+2d_1+2d_2+c_1c_2+38)\,\mathbf{u}^2\right.
\end{aligned}
$$

$$
\begin{aligned}
&+ (8c_1 + 8c_2 + 6d_1 + 6d_2 + 125)\,\mathbf{u}^3 \\
&+ 44\mathbf{u}^4 - 54\mathbf{u}^5 - 92\mathbf{u}^6 - 45\mathbf{u}^7 + 7\mathbf{u}^8 + 17\mathbf{u}^9 + 7\mathbf{u}^{10} + \mathbf{u}^{11})\Big)\mathrm{fl}\,(f_1 \cdot f_2) \\
&+ (Y + X/2)\,\mathbf{u}_S \\
\leq &\Big(\mathrm{float}\,(c_1 + c_2 + 1)\,\mathbf{u} + \frac{1}{(1+\mathbf{u})^6}\,\big((10c_1 + 10c_2 + 2d_1 + 2d_2 + c_1c_2 + 38)\,\mathbf{u}^2 \\
&+ (8U + 8U + 6U + 6U + 125)\,\mathbf{u}^3 \\
&+ 44\mathbf{u}^4 - 54\mathbf{u}^5 - 92\mathbf{u}^6 - 45\mathbf{u}^7 + 7\mathbf{u}^8 + 17\mathbf{u}^9 + 7\mathbf{u}^{10} + \mathbf{u}^{11})\Big)\mathrm{fl}\,(f_1 \cdot f_2) \\
&+ (Y + X/2)\,\mathbf{u}_S \\
= &\Big(\mathrm{float}\,(c_1 + c_2 + 1)\,\mathbf{u} + \frac{1}{(1+\mathbf{u})^6}\,\big((10c_1 + 10c_2 + 2d_1 + 2d_2 + c_1c_2 + 66)\,\mathbf{u}^2 \\
&+ 97\mathbf{u}^3 + 44\mathbf{u}^4 - 54\mathbf{u}^5 - 92\mathbf{u}^6 - 45\mathbf{u}^7 + 7\mathbf{u}^8 + 17\mathbf{u}^9 + 7\mathbf{u}^{10} + \mathbf{u}^{11})\Big)\mathrm{fl}\,(f_1 \cdot f_2) \\
&+ (Y + X/2)\,\mathbf{u}_S \\
< &\left(\mathrm{float}\,(c_1 + c_2 + 1)\,\mathbf{u} + \frac{1}{(1+\mathbf{u})^6}\,\big((10c_1 + 10c_2 + 2d_1 + 2d_2 + c_1c_2 + 67)\,\mathbf{u}^2\big)\right)\mathrm{fl}\,(f_1 \cdot f_2) \\
&+ (Y + X/2)\,\mathbf{u}_S \\
\leq &\Big(\mathrm{float}\,(c_1 + c_2 + 1)\,\mathbf{u} + \frac{1}{(1+\mathbf{u})^6}\,\big(\big((1+\mathbf{u})\mathrm{fl}(10c_1) + (1+\mathbf{u})\mathrm{fl}(10c_2) \\
&+ \mathrm{fl}(2d_1) + \mathrm{fl}(2d_2) + (1+\mathbf{u})\mathrm{fl}(c_1 \cdot c_2) + 67\big)\mathbf{u}^2\big)\Big)\mathrm{fl}\,(f_1 \cdot f_2) \\
&+ (Y + X/2)\,\mathbf{u}_S \\
< &\Big(\mathrm{float}\,(c_1 + c_2 + 1)\,\mathbf{u} + \frac{1}{(1+\mathbf{u})^5}\,\big(\big(\mathrm{fl}(10c_1) + \mathrm{fl}(10c_2) \\
&+ \mathrm{fl}(2d_1) + \mathrm{fl}(2d_2) + \mathrm{fl}(c_1 \cdot c_2) + 67\big)\mathbf{u}^2\big)\Big)\mathrm{fl}\,(f_1 \cdot f_2) \\
&+ (Y + X/2)\,\mathbf{u}_S \\
\leq &\Big(\mathrm{float}\,(c_1 + c_2 + 1)\,\mathbf{u} \\
&+ \frac{1}{(1+\mathbf{u})^5}(1+\mathbf{u})^5\mathrm{float}\,(10c_1 + 10c_2 + 2d_1 + 2d_2 + c_1 \cdot c_2 + 67)\,\mathbf{u}^2\Big)\mathrm{fl}\,(f_1 \cdot f_2) \\
&+ (Y + X/2)\,\mathbf{u}_S \\
= &\Big(\mathrm{float}\,(c_1 + c_2 + 1)\,\mathbf{u} + \mathrm{float}\,(10c_1 + 10c_2 + 2d_1 + 2d_2 + c_1 \cdot c_2 + 67)\,\mathbf{u}^2\Big)\mathrm{fl}\,(f_1 \cdot f_2) \\
&+ (Y + X/2)\,\mathbf{u}_S \\
= &\Big(\mathrm{float}\,(c_1 + c_2 + 1)\,\mathbf{u} + \mathrm{float}\,(10c_1 + 10c_2 + 2d_1 + 2d_2 + c_1 \cdot c_2 + 67)\,\mathbf{u}^2\Big)\mathrm{fl}\,(f_1 \cdot f_2) \\
&+ \Big(\big(1 + c_1\mathbf{u} + d_1\mathbf{u}^2\big)\,f_1e_2 + \big(1 + c_2\mathbf{u} + d_2\mathbf{u}^2\big)\,f_2e_1 + \frac{1}{2} + e_1e_2\mathbf{u}_S \\
&+ \big((c_1 + c_2)\,\mathbf{u} + (c_1c_2 + d_1 + d_2)\,\mathbf{u}^2 + (c_1d_2 + c_2d_1)\,\mathbf{u}^3 + d_1d_2\mathbf{u}^4\big)/2\Big)\mathbf{u}_S
\end{aligned}
$$

$$
\begin{aligned}
&\leq \Big(\text{float}\left(c_1+c_2+1\right)\mathbf{u}+\text{float}\left(10c_1+10c_2+2d_1+2d_2+c_1{\cdot}c_2+67\right)\mathbf{u}^2\Big)\,\text{fl}\left(f_1{\cdot}f_2\right)\\
&\quad+\Big(\left(1+U\mathbf{u}+U\mathbf{u}^2\right)f_1e_2+\left(1+U\mathbf{u}+U\mathbf{u}^2\right)f_2e_1+\frac{1}{2}+e_1e_2\mathbf{u}_S\\
&\quad+\left(\left(U+U\right)\mathbf{u}+\left(U^2+U+U\right)\mathbf{u}^2+\left(U^2+U^2\right)\mathbf{u}^3+U^2\mathbf{u}^4\right)/2\Big)\mathbf{u}_S\\
&=\Big(\text{float}\left(c_1+c_2+1\right)\mathbf{u}+\text{float}\left(10c_1+10c_2+2d_1+2d_2+c_1{\cdot}c_2+67\right)\mathbf{u}^2\Big)\,\text{fl}\left(f_1{\cdot}f_2\right)\\
&\quad+\left(\left(2-\mathbf{u}^2\right)f_1e_2+\left(2-\mathbf{u}^2\right)f_2e_1+\frac{1}{2}\left(4-4\mathbf{u}^2+\mathbf{u}^4\right)+e_1e_2\mathbf{u}_S\right)\mathbf{u}_S\\
&=\big(\text{float}\left(c_1+c_2+1\right)\mathbf{u}+\text{float}\left(10c_1+10c_2+2d_1+2d_2+c_1{\cdot}c_2+67\right)\mathbf{u}^2\big)\,\text{fl}\left(f_1{\cdot}f_2\right)\\
&\quad+\frac{(1+\mathbf{u})^5}{(1+\mathbf{u})^5}\left(\left(2-\mathbf{u}^2\right)f_1e_2+\left(2-\mathbf{u}^2\right)f_2e_1+\frac{1}{2}\left(4-4\mathbf{u}^2+\mathbf{u}^4\right)+e_1e_2\mathbf{u}_S\right)\mathbf{u}_S\\
&=\big(\text{float}\left(c_1+c_2+1\right)\mathbf{u}+\text{float}\left(10c_1+10c_2+2d_1+2d_2+c_1{\cdot}c_2+67\right)\mathbf{u}^2\big)\,\text{fl}\left(f_1{\cdot}f_2\right)\\
&\quad+\frac{1}{(1+\mathbf{u})^5}\Big((1+\mathbf{u})^5\left(2-\mathbf{u}^2\right)f_1e_2+(1+\mathbf{u})^5\left(2-\mathbf{u}^2\right)f_2e_1\\
&\quad+\frac{(1+\mathbf{u})^5}{2}\left(4-4\mathbf{u}^2+\mathbf{u}^4\right)+(1+\mathbf{u})^5e_1e_2\mathbf{u}_S\Big)\mathbf{u}_S\\
&\leq\big(\text{float}\left(c_1+c_2+1\right)\mathbf{u}+\text{float}\left(10c_1+10c_2+2d_1+2d_2+c_1{\cdot}c_2+67\right)\mathbf{u}^2\big)\,\text{fl}\left(f_1{\cdot}f_2\right)\\
&\quad+\frac{1}{(1+\mathbf{u})^5}\Big((1+\mathbf{u})^5\left(2-\mathbf{u}^2\right)\left((1+\mathbf{u})\text{fl}(f_1{\cdot}e_2)+\frac{\mathbf{u}_S}{2}\right)\\
&\quad+(1+\mathbf{u})^5\left(2-\mathbf{u}^2\right)\left((1+\mathbf{u})\text{fl}(f_2{\cdot}e_1)+\frac{\mathbf{u}_S}{2}\right)\\
&\quad+\frac{(1+\mathbf{u})^5}{2}\left(4-4\mathbf{u}^2+\mathbf{u}^4\right)+(1+\mathbf{u})^5(1+\mathbf{u})\text{fl}(e_1{\cdot}e_2)\mathbf{u}_S\Big)\mathbf{u}_S\\
&=\big(\text{float}\left(c_1+c_2+1\right)\mathbf{u}+\text{float}\left(10c_1+10c_2+2d_1+2d_2+c_1{\cdot}c_2+67\right)\mathbf{u}^2\big)\,\text{fl}\left(f_1{\cdot}f_2\right)\\
&\quad+\frac{1}{(1+\mathbf{u})^4}\Big((1+\mathbf{u})^5\left(2-\mathbf{u}^2\right)\text{fl}(f_1{\cdot}e_2)+(1+\mathbf{u})^5\left(2-\mathbf{u}^2\right)\text{fl}(f_2{\cdot}e_1)\\
&\quad+\frac{(1+\mathbf{u})^4}{2}\left(4-4\mathbf{u}^2+\mathbf{u}^4+(4-2\mathbf{u}^2)\mathbf{u}_S\right)+(1+\mathbf{u})^5\text{fl}(e_1{\cdot}e_2)\mathbf{u}_S\Big)\mathbf{u}_S\\
&\leq\left(\text{float}\left(c_1+c_2+1\right)\mathbf{u}+\text{float}\left(10c_1+10c_2+2d_1+2d_2+c_1{\cdot}c_2+67\right)\mathbf{u}^2\right)\text{fl}\left(f_1{\cdot}f_2\right)\\
&\quad+\frac{1}{(1+\mathbf{u})^4}\Big((1+\mathbf{u})^5\left(2-\mathbf{u}^2\right)\text{fl}(f_1{\cdot}e_2)+(1+\mathbf{u})^5\left(2-\mathbf{u}^2\right)\text{fl}(f_2{\cdot}e_1)\\
&\quad+\frac{(1+\mathbf{u})^4}{2}\left(4-4\mathbf{u}^2+\mathbf{u}^4+(4-2\mathbf{u}^2)\mathbf{u}_S\right)+(1+\mathbf{u})^5\left(\text{fl}((e_1{\cdot}e_2)\mathbf{u}_S)+\frac{\mathbf{u}_S}{2}\right)\Big)\mathbf{u}_S\\
&=\left(\text{float}\left(c_1+c_2+1\right)\mathbf{u}+\text{float}\left(10c_1+10c_2+2d_1+2d_2+c_1{\cdot}c_2+67\right)\mathbf{u}^2\right)\text{fl}\left(f_1{\cdot}f_2\right)\\
&\quad+\frac{1}{(1+\mathbf{u})^4}\Big((1+\mathbf{u})^5\left(2-\mathbf{u}^2\right)\text{fl}(f_1{\cdot}e_2)+(1+\mathbf{u})^5\left(2-\mathbf{u}^2\right)\text{fl}(f_2{\cdot}e_1)\\
&\quad+\frac{(1+\mathbf{u})^4}{2}\left(4-4\mathbf{u}^2+\mathbf{u}^4+(5+\mathbf{u}-2\mathbf{u}^2)\mathbf{u}_S\right)+(1+\mathbf{u})^5\text{fl}((e_1{\cdot}e_2)\mathbf{u}_S)\Big)\mathbf{u}_S\\
&<\left(\text{float}\left(c_1+c_2+1\right)\mathbf{u}+\text{float}\left(10c_1+10c_2+2d_1+2d_2+c_1{\cdot}c_2+67\right)\mathbf{u}^2\right)\text{fl}\left(f_1{\cdot}f_2\right)\\
&\quad+\frac{1}{(1+\mathbf{u})^4}\Big(\text{fl}(2+12\mathbf{u})\text{fl}(f_1{\cdot}e_2)+\text{fl}(2+12\mathbf{u})\text{fl}(f_2{\cdot}e_1)
\end{aligned}
$$

$$
\begin{aligned}
&\quad + \frac{(1+\mathbf{u})^4}{2}\left(4 - 4\mathbf{u}^2 + \mathbf{u}^4 + (5 + \mathbf{u} - 2\mathbf{u}^2)\mathbf{u}_S\right) + \mathrm{fl}(1+6\mathbf{u})\mathrm{fl}((e_1 \cdot e_2)\mathbf{u}_S)\Big)\mathbf{u}_S \\
&\leq \left(\mathrm{float}\,(c_1 + c_2 + 1)\,\mathbf{u} + \mathrm{float}\,(10c_1 + 10c_2 + 2d_1 + 2d_2 + c_1 \cdot c_2 + 67)\,\mathbf{u}^2\right)\mathrm{fl}\,(f_1 \cdot f_2) \\
&\quad + \frac{1}{(1+\mathbf{u})^4}\Bigg(\Big((1+\mathbf{u})\mathrm{fl}\big((2+12\mathbf{u})\cdot(f_1 \cdot e_2)\big) + \frac{\mathbf{u}_S}{2}\Big) \\
&\quad + \Big((1+\mathbf{u})\mathrm{fl}\big((2+12\mathbf{u})\cdot(f_2 \cdot e_1)\big) + \frac{\mathbf{u}_S}{2}\Big) \\
&\quad + \frac{(1+\mathbf{u})^4}{2}\left(4 - 4\mathbf{u}^2 + \mathbf{u}^4 + (5 + \mathbf{u} - 2\mathbf{u}^2)\mathbf{u}_S\right) \\
&\quad + \Big((1+\mathbf{u})\mathrm{fl}((1+6\mathbf{u})\cdot((e_1 \cdot e_2)\mathbf{u}_S)) + \frac{\mathbf{u}_S}{2}\Big)\Bigg)\mathbf{u}_S \\
&= \left(\mathrm{float}\,(c_1 + c_2 + 1)\,\mathbf{u} + \mathrm{float}\,(10c_1 + 10c_2 + 2d_1 + 2d_2 + c_1 \cdot c_2 + 67)\,\mathbf{u}^2\right)\mathrm{fl}\,(f_1 \cdot f_2) \\
&\quad + \frac{1}{(1+\mathbf{u})^4}\Bigg((1+\mathbf{u})\mathrm{fl}\big((2+12\mathbf{u})\cdot(f_1 \cdot e_2)\big) + (1+\mathbf{u})\mathrm{fl}\big((2+12\mathbf{u})\cdot(f_2 \cdot e_1)\big) \\
&\quad + \frac{(1+\mathbf{u})^4}{2}\left(4 - 4\mathbf{u}^2 + \mathbf{u}^4 + \left(5 + \mathbf{u} - 2\mathbf{u}^2 + \frac{3}{(1+\mathbf{u})^4}\right)\mathbf{u}_S\right) \\
&\quad + (1+\mathbf{u})\mathrm{fl}\big((1+6\mathbf{u})\cdot((e_1 \cdot e_2)\mathbf{u}_S)\big)\Bigg)\mathbf{u}_S \\
&= \left(\mathrm{float}\,(c_1 + c_2 + 1)\,\mathbf{u} + \mathrm{float}\,(10c_1 + 10c_2 + 2d_1 + 2d_2 + c_1 \cdot c_2 + 67)\,\mathbf{u}^2\right)\mathrm{fl}\,(f_1 \cdot f_2) \\
&\quad + \frac{1}{(1+\mathbf{u})^3}\Bigg(\mathrm{fl}((2+12\mathbf{u})\cdot(f_1 \cdot e_2)) + \mathrm{fl}((2+12\mathbf{u})\cdot(f_2 \cdot e_1)) \\
&\quad + \frac{(1+\mathbf{u})^3}{2}\left(4 - 4\mathbf{u}^2 + \mathbf{u}^4 + \left(5 + \mathbf{u} - 2\mathbf{u}^2 + \frac{3}{(1+\mathbf{u})^4}\right)\mathbf{u}_S\right) \\
&\quad + \mathrm{fl}\big((1+6\mathbf{u})\cdot((e_1 \cdot e_2)\mathbf{u}_S)\big)\Bigg)\mathbf{u}_S \\
&< \left(\mathrm{float}\,(c_1 + c_2 + 1)\,\mathbf{u} + \mathrm{float}\,(10c_1 + 10c_2 + 2d_1 + 2d_2 + c_1 \cdot c_2 + 67)\,\mathbf{u}^2\right)\mathrm{fl}\,(f_1 \cdot f_2) \\
&\quad + \frac{(\mathrm{fl}((2+12\mathbf{u})\cdot(f_1 \cdot e_2)) + \mathrm{fl}((2+12\mathbf{u})\cdot(f_2 \cdot e_1)) + 3 + \mathrm{fl}((1+6\mathbf{u})\cdot((e_1 \cdot e_2)\mathbf{u}_S)))\,\mathbf{u}_S}{(1+\mathbf{u})^3} \\
&\leq \left(\mathrm{float}\,(c_1 + c_2 + 1)\,\mathbf{u} + \mathrm{float}\,(10c_1 + 10c_2 + 2d_1 + 2d_2 + c_1 \cdot c_2 + 67)\,\mathbf{u}^2\right)\mathrm{fl}\,(f_1 \cdot f_2) \\
&\quad + \frac{(1+\mathbf{u})^3\mathrm{float}\,((2+12\mathbf{u})\cdot(f_1 \cdot e_2) + (2+12\mathbf{u})\cdot(f_2 \cdot e_1) + 3 + (1+6\mathbf{u})\cdot((e_1 \cdot e_2)\mathbf{u}_S))\,\mathbf{u}_S}{(1+\mathbf{u})^3} \\
&= \left(\mathrm{fl}\,(c_1 + c_2 + 1)\,\mathbf{u} + \mathrm{float}\,(10c_1 + 10c_2 + 2d_1 + 2d_2 + c_1c_2 + 67)\,\mathbf{u}^2\right)\mathrm{float}\,(f_1 \cdot f_2) \\
&\quad + \mathrm{float}\,((2+12\mathbf{u})\cdot(f_1 \cdot e_2) + (2+12\mathbf{u})\cdot(f_2 \cdot e_1) + 3 + (1+6\mathbf{u})\cdot((e_1 \cdot e_2)\mathbf{u}_S))\,\mathbf{u}_S \\
&\leq \left(\mathrm{fl}\,(c_1 + c_2 + 1)\,\mathbf{u} + \mathrm{float}\,(10c_1 + 10c_2 + 2d_1 + 2d_2 + c_1c_2 + 67)\,\mathbf{u}^2\right)\mathrm{float}\,(f_1 \cdot f_2) \\
&\quad + \lceil \mathrm{float}\,((2+12\mathbf{u})\cdot(f_1 \cdot e_2) + (2+12\mathbf{u})\cdot(f_2 \cdot e_1) + 3 + (1+6\mathbf{u})\cdot((e_1 \cdot e_2)\mathbf{u}_S))\rceil\,\mathbf{u}_S
\end{aligned}
$$

□

定理 5.10 の証明

定理 5.9 の証明と同様に U を用いて誤差評価を行う．また, X, Y を次のようにおく．

$$X = (2c_1 + 2c_2)\,\mathbf{u} + (c_1c_2 + 2d_1 + 2d_2)\,\mathbf{u}^2 + (c_1d_2 + c_2d_1)\,\mathbf{u}^3 + d_1d_2\mathbf{u}^4,$$
$$Y = \left(2 + c_1\mathbf{u} + d_1\mathbf{u}^2\right) f_1e_2 + \left(2 + c_2\mathbf{u} + d_2\mathbf{u}^2\right) f_2e_1 + \frac{1}{2} + e_1e_2\mathbf{u}_S.$$

これを利用して, $\mathrm{fl}(x\cdot y)$, ab に対する誤差上限を評価する．ただし, $f_1 = \mathrm{ufp}(f_1)$, $f_2 = \mathrm{ufp}(f_2)$ より, これらとの積における丸め誤差はアンダーフローのみを考えればよい．また, float で評価する際は補題 2.6 を利用する．

$$\begin{aligned}
&|\mathrm{fl}(x\cdot y) - (ab)| \\
=&|\mathrm{fl}(x\cdot y) - xy + xy - ab| \\
\le&|\mathrm{fl}(x\cdot y) - xy| + |xy - ab| \\
=&|\mathrm{fl}(x\cdot y) - xy| + |xy - (x+\delta_1)(y+\delta_2)| \\
=&|\mathrm{fl}(x\cdot y) - xy| + |-x\delta_2 - \delta_1 y - \delta_1\delta_2| \\
\le&\mathbf{u}\mathrm{fl}(|x\cdot y|) + \frac{\mathbf{u}_S}{2} + |x||\delta_2| + |\delta_1||y| + |\delta_1||\delta_2| \\
\le&\mathbf{u}\mathrm{fl}(2f_1 2f_2) + \frac{\mathbf{u}_S}{2} + 2f_1\left(\left(c_2\mathbf{u} + d_2\mathbf{u}^2\right) f_2 + e_2\mathbf{u}_S\right) + \left(\left(c_1\mathbf{u} + d_1\mathbf{u}^2\right) f_1 + e_1\mathbf{u}_S\right) 2f_2 \\
&+ \left(\left(c_1\mathbf{u} + d_1\mathbf{u}^2\right) f_1 + e_1\mathbf{u}_S\right)\left(\left(c_2\mathbf{u} + d_2\mathbf{u}^2\right) f_2 + e_2\mathbf{u}_S\right) \\
=&\mathbf{u}\mathrm{fl}(4f_1\cdot f_2) + \frac{\mathbf{u}_S}{2} + 2\left(c_2\mathbf{u} + d_2\mathbf{u}^2\right) f_1f_2 + 2f_1e_2\mathbf{u}_S + 2\left(c_1\mathbf{u} + d_1\mathbf{u}^2\right) f_1f_2 + 2f_2e_1\mathbf{u}_S \\
&+ \left(c_1\mathbf{u} + d_1\mathbf{u}^2\right)\left(c_2\mathbf{u} + d_2\mathbf{u}^2\right) f_1f_2 + \left(c_1\mathbf{u} + d_1\mathbf{u}^2\right) f_1e_2\mathbf{u}_S + \left(c_2\mathbf{u} + d_2\mathbf{u}^2\right) e_1f_2\mathbf{u}_S \\
&+ e_1e_2\mathbf{u}_S^2 \\
=&4\mathbf{u}\mathrm{fl}(f_1\cdot f_2) \\
&+ \left((2c_1 + 2c_2)\,\mathbf{u} + (c_1c_2 + 2d_1 + 2d_2)\,\mathbf{u}^2 + (c_1d_2 + c_2d_1)\,\mathbf{u}^3 + d_1d_2\mathbf{u}^4\right) f_1f_2 \\
&+ \left(\left(2 + c_1\mathbf{u} + d_1\mathbf{u}^2\right) f_1e_2 + \left(2 + c_2\mathbf{u} + d_2\mathbf{u}^2\right) f_2e_1 + \frac{1}{2} + e_1e_2\mathbf{u}_S\right)\mathbf{u}_S \\
=&4\mathbf{u}\mathrm{fl}(f_1\cdot f_2) + Xf_1f_2 + Y\mathbf{u}_S \\
\le&4\mathbf{u}\mathrm{fl}(f_1\cdot f_2) + X\left(\mathrm{fl}\left(f_1\cdot f_2\right) + \frac{\mathbf{u}_S}{2}\right) + Y\mathbf{u}_S \\
=&\left(X + 4\mathbf{u}\right)\mathrm{fl}\left(f_1\cdot f_2\right) + \left(Y + X/2\right)\mathbf{u}_S \\
=&\left((2c_1 + 2c_2 + 4)\,\mathbf{u} + (2d_1 + 2d_2 + c_1c_2)\,\mathbf{u}^2 + (c_1d_2 + c_2d_1)\,\mathbf{u}^3 + d_1d_2\mathbf{u}^4\right)\mathrm{fl}\left(f_1\cdot f_2\right) \\
&+ \left(Y + X/2\right)\mathbf{u}_S \\
=&\left(\frac{(1+\mathbf{u})^2}{(1+\mathbf{u})^2}\left(2c_1 + 2c_2 + 4\right)\mathbf{u} + \left(2d_1 + 2d_2 + c_1c_2\right)\mathbf{u}^2\right. \\
&\left. + \left(c_1d_2 + c_2d_1\right)\mathbf{u}^3 + d_1d_2\mathbf{u}^4\right)\mathrm{fl}\left(f_1\cdot f_2\right) \\
&+ \left(Y + X/2\right)\mathbf{u}_S \\
=&\left(\frac{1}{(1+\mathbf{u})^2}\left(2c_1 + 2c_2 + 4\right)\mathbf{u} + \left(2d_1 + 2d_2 + c_1c_2 + 2\frac{2c_1 + 2c_2 + 4}{(1+\mathbf{u})^2}\right)\mathbf{u}^2\right.
\end{aligned}$$

$$
\begin{aligned}
&+\left(c_1d_2+c_2d_1+\frac{2c_1+2c_2+4}{(1+\mathbf{u})^2}\right)\mathbf{u}^3+d_1d_2\mathbf{u}^4\Big)\mathrm{fl}\left(f_1\cdot f_2\right)\\
&+\left(Y+X/2\right)\mathbf{u}_S\\
=&\Bigg(\frac{1}{(1+\mathbf{u})^2}(1+\mathbf{u})^2\mathrm{float}\left(2c_1+2c_2+4\right)\mathbf{u}+\left(2d_1+2d_2+c_1c_2+2\frac{2c_1+2c_2+4}{(1+\mathbf{u})^2}\right)\mathbf{u}^2\\
&+\left(c_1d_2+c_2d_1+\frac{2c_1+2c_2+4}{(1+\mathbf{u})^2}\right)\mathbf{u}^3+d_1d_2\mathbf{u}^4\Big)\mathrm{fl}\left(f_1\cdot f_2\right)\\
&+\left(Y+X/2\right)\mathbf{u}_S\\
=&\Bigg(\mathrm{float}\left(2c_1+2c_2+4\right)\mathbf{u}+\left(2d_1+2d_2+c_1c_2+2\frac{2c_1+2c_2+4}{(1+\mathbf{u})^2}\right)\mathbf{u}^2\\
&+\left(c_1d_2+c_2d_1+\frac{2c_1+2c_2+4}{(1+\mathbf{u})^2}\right)\mathbf{u}^3+d_1d_2\mathbf{u}^4\Big)\mathrm{fl}\left(f_1\cdot f_2\right)\\
&+\left(Y+X/2\right)\mathbf{u}_S\\
=&\Bigg(\mathrm{float}\left(2c_1+2c_2+4\right)\mathbf{u}+\frac{(1+\mathbf{u})^6}{(1+\mathbf{u})^6}\left(2d_1+2d_2+c_1c_2+2\frac{2c_1+2c_2+4}{(1+\mathbf{u})^2}\right)\mathbf{u}^2\\
&+\frac{(1+\mathbf{u})^6}{(1+\mathbf{u})^6}\left(c_1d_2+c_2d_1+\frac{2c_1+2c_2+4}{(1+\mathbf{u})^2}\right)\mathbf{u}^3+\frac{(1+\mathbf{u})^6}{(1+\mathbf{u})^6}d_1d_2\mathbf{u}^4\Big)\mathrm{fl}\left(f_1\cdot f_2\right)\\
&+\left(Y+X/2\right)\mathbf{u}_S\\
=&\Bigg(\mathrm{float}\left(2c_1+2c_2+4\right)\mathbf{u}+\frac{1}{(1+\mathbf{u})^6}\Bigg(\left(4c_1+4c_2+2d_1+2d_2+c_1c_2+8\right)\mathbf{u}^2\\
&+\left(18c_1+18c_2+12d_1+12d_2+6c_1c_2+c_1d_2+c_2d_1+36\right)\mathbf{u}^3\\
&+\left(32c_1+32c_2+30d_1+30d_2+15c_1c_2+6c_1d_2+6c_2d_1+d_1d_2+64\right)\mathbf{u}^4\\
&+\left(28c_1+28c_2+40d_1+40d_2+20c_1c_2+15c_1d_2+15c_2d_1+6d_1d_2+56\right)\mathbf{u}^5\\
&+\left(12c_1+12c_2+30d_1+30d_2+15c_1c_2+20c_1d_2+20c_2d_1+15d_1d_2+24\right)\mathbf{u}^6\\
&+\left(2c_1+2c_2+12d_1+12d_2+6c_1c_2+15c_1d_2+15c_2d_1+20d_1d_2+4\right)\mathbf{u}^7\\
&+\left(2d_1+2d_2+c_1c_2+6c_1d_2+6c_2d_1+15d_1d_2\right)\mathbf{u}^8\\
&+\left(c_1d_2+c_2d_1+6d_1d_2\right)\mathbf{u}^9+d_1d_2\mathbf{u}^{10}\Bigg)\mathrm{fl}\left(f_1\cdot f_2\right)\\
&+\left(Y+X/2\right)\mathbf{u}_S\\
\leq&\Bigg(\mathrm{float}\left(2c_1+2c_2+4\right)\mathbf{u}+\frac{1}{(1+\mathbf{u})^6}\Bigg(\left(4c_1+4c_2+2d_1+2d_2+c_1c_2+8\right)\mathbf{u}^2\\
&+\left(18c_1+18c_2+12d_1+12d_2+6c_1c_2+c_1d_2+c_2d_1+36\right)\mathbf{u}^3\\
&+\left(32U+32U+30U+30U+15U^2+6U^2+6U^2+U^2+64\right)\mathbf{u}^4\\
&+\left(28U+28U+40U+40U+20U^2+15U^2+15U^2+6U^2+56\right)\mathbf{u}^5\\
&+\left(12U+12U+30U+30U+15U^2+20U^2+20U^2+15U^2+24\right)\mathbf{u}^6\\
&+\left(2U+2U+12U+12U+6U^2+15U^2+15U^2+20U^2+4\right)\mathbf{u}^7\\
&+\left(2U+2U+U^2+6U^2+6U^2+15U^2\right)\mathbf{u}^8
\end{aligned}
$$

$$
\begin{aligned}
&+\left(U^2+U^2+6U^2\right)\mathbf{u}^9+U^2\mathbf{u}^{10}\Big)\,\mathrm{fl}\left(f_1\cdot f_2\right)\\
&+\left(Y+X/2\right)\mathbf{u}_S\\
=&\Big(\mathrm{float}\left(2c_1+2c_2+4\right)\mathbf{u}+\frac{1}{(1+\mathbf{u})^6}\Big(\left(4c_1+4c_2+2d_1+2d_2+c_1c_2+36\right)\mathbf{u}^2\\
&+\left(18c_1+18c_2+12d_1+12d_2+6c_1c_2+c_1d_2+c_2d_1+160\right)\mathbf{u}^3\\
&+62\mathbf{u}^4-24\mathbf{u}^5-46\mathbf{u}^6-12\mathbf{u}^7+9\mathbf{u}^8+6\mathbf{u}^9+\mathbf{u}^{10}\Big)\,\mathrm{fl}\left(f_1\cdot f_2\right)\\
&+\left(Y+X/2\right)\mathbf{u}_S\\
\le&\Big(\mathrm{float}\left(2c_1+2c_2+4\right)\mathbf{u}+\frac{1}{(1+\mathbf{u})^6}\Big(\left(4c_1+4c_2+2d_1+2d_2+c_1c_2+36\right)\mathbf{u}^2\\
&+\left(18U+18U+12U+12U+6(c_1U+Uc_2)+Ud_2+Ud_1+160\right)\mathbf{u}^3\\
&+62\mathbf{u}^4-24\mathbf{u}^5-46\mathbf{u}^6-12\mathbf{u}^7+9\mathbf{u}^8+6\mathbf{u}^9+\mathbf{u}^{10}\Big)\,\mathrm{fl}\left(f_1\cdot f_2\right)\\
&+\left(Y+X/2\right)\mathbf{u}_S\\
=&\Big(\mathrm{float}\left(2c_1+2c_2+4\right)\mathbf{u}+\frac{1}{(1+\mathbf{u})^6}\Big(\left(10c_1+10c_2+3d_1+3d_2+c_1c_2+96\right)\mathbf{u}^2\\
&+\left(-6c_1-6c_2-d_1-d_2+100\right)\mathbf{u}^3\\
&+62\mathbf{u}^4-24\mathbf{u}^5-46\mathbf{u}^6-12\mathbf{u}^7+9\mathbf{u}^8+6\mathbf{u}^9+\mathbf{u}^{10}\Big)\,\mathrm{fl}\left(f_1\cdot f_2\right)\\
&+\left(Y+X/2\right)\mathbf{u}_S\\
<&\Big(\mathrm{float}\left(2c_1+2c_2+4\right)\mathbf{u}+\frac{1}{(1+\mathbf{u})^6}\left(10c_1+10c_2+3d_1+3d_2+c_1c_2+97\right)\mathbf{u}^2\Big)\,\mathrm{fl}\left(f_1\cdot f_2\right)\\
&+\left(Y+X/2\right)\mathbf{u}_S\\
\le&\Big(\mathrm{float}\left(2c_1+2c_2+4\right)\mathbf{u}+\frac{1}{(1+\mathbf{u})^6}\Big((1+\mathbf{u})\mathrm{fl}\left(10c_1\right)+(1+\mathbf{u})\mathrm{fl}\left(10c_2\right)\\
&+(1+\mathbf{u})\mathrm{fl}\left(3d_1\right)+(1+\mathbf{u})\mathrm{fl}\left(3d_2\right)+(1+\mathbf{u})\mathrm{fl}\left(c_1\cdot c_2\right)+97\Big)\mathbf{u}^2\Big)\,\mathrm{fl}\left(f_1\cdot f_2\right)\\
&+\left(Y+X/2\right)\mathbf{u}_S\\
=&\Big(\mathrm{float}\left(2c_1+2c_2+4\right)\mathbf{u}+\frac{1}{(1+\mathbf{u})^5}\Big(\mathrm{fl}\left(10c_1\right)+\mathrm{fl}\left(10c_2\right)\\
&+\mathrm{fl}\left(3d_1\right)+\mathrm{fl}\left(3d_2\right)+\mathrm{fl}\left(c_1\cdot c_2\right)+97\Big)\mathbf{u}^2\Big)\,\mathrm{fl}\left(f_1\cdot f_2\right)\\
&+\left(Y+X/2\right)\mathbf{u}_S\\
\le&\Big(\mathrm{float}\left(2c_1+2c_2+4\right)\mathbf{u}\\
&+\frac{1}{(1+\mathbf{u})^5}(1+\mathbf{u})^5\mathrm{float}\Big(\left(10c_1\right)+\left(10c_2\right)+\left(3d_1\right)+\left(3d_2\right)+\left(c_1\cdot c_2\right)+97\Big)\mathbf{u}^2\Big)\,\mathrm{fl}\left(f_1\cdot f_2\right)\\
&+\left(Y+X/2\right)\mathbf{u}_S
\end{aligned}
$$

$$
\begin{aligned}
=&\Big(\text{float}\left(2c_1+2c_2+4\right)\mathbf{u}+\text{float}\left(10c_1+10c_2+3d_1+3d_2+c_1{\cdot}c_2+97\right)\mathbf{u}^2\Big)\,\text{fl}\left(f_1{\cdot}f_2\right)\\
&+\left(Y+X/2\right)\mathbf{u}_S\\
=&\Big(\text{float}\left(2c_1+2c_2+4\right)\mathbf{u}+\text{float}\left(10c_1+10c_2+3d_1+3d_2+c_1{\cdot}c_2+97\right)\mathbf{u}^2\Big)\,\text{fl}\left(f_1{\cdot}f_2\right)\\
&+\Bigg(\left(2+c_1\mathbf{u}+d_1\mathbf{u}^2\right)f_1e_2+\left(2+c_2\mathbf{u}+d_2\mathbf{u}^2\right)f_2e_1+\frac{1}{2}+e_1e_2\mathbf{u}_S\\
&+\Big(\left(2c_1+2c_2\right)\mathbf{u}+\left(c_1c_2+2d_1+2d_2\right)\mathbf{u}^2+\left(c_1d_2+c_2d_1\right)\mathbf{u}^3+d_1d_2\mathbf{u}^4\Big)/2\Bigg)\mathbf{u}_S\\
\le&\Big(\text{float}\left(2c_1+2c_2+4\right)\mathbf{u}+\text{float}\left(10c_1+10c_2+3d_1+3d_2+c_1{\cdot}c_2+97\right)\mathbf{u}^2\Big)\,\text{fl}\left(f_1{\cdot}f_2\right)\\
&+\Bigg(\left(2+U\mathbf{u}+U\mathbf{u}^2\right)f_1e_2+\left(2+U\mathbf{u}+U\mathbf{u}^2\right)f_2e_1+\frac{1}{2}+e_1e_2\mathbf{u}_S\\
&+\Big(\left(2U+2U\right)\mathbf{u}+\left(U^2+2U+2U\right)\mathbf{u}^2+\left(U^2+U^2\right)\mathbf{u}^3+U^2\mathbf{u}^4\Big)/2\Bigg)\mathbf{u}_S\\
=&\Big(\text{float}\left(2c_1+2c_2+4\right)\mathbf{u}+\text{float}\left(10c_1+10c_2+3d_1+3d_2+c_1{\cdot}c_2+97\right)\mathbf{u}^2\Big)\,\text{fl}\left(f_1{\cdot}f_2\right)\\
&+\Bigg(\left(3-\mathbf{u}^2\right)f_1e_2+\left(3-\mathbf{u}^2\right)f_2e_1+\frac{1}{2}\left(6-6\mathbf{u}^2+\mathbf{u}^4\right)+e_1e_2\mathbf{u}_S\Bigg)\mathbf{u}_S\\
=&\Big(\text{float}\left(2c_1+2c_2+4\right)\mathbf{u}+\text{float}\left(10c_1+10c_2+3d_1+3d_2+c_1{\cdot}c_2+97\right)\mathbf{u}^2\Big)\,\text{fl}\left(f_1{\cdot}f_2\right)\\
&+\frac{(1+\mathbf{u})^4}{(1+\mathbf{u})^4}\Bigg(\left(3-\mathbf{u}^2\right)f_1e_2+\left(3-\mathbf{u}^2\right)f_2e_1+\frac{1}{2}\left(6-6\mathbf{u}^2+\mathbf{u}^4\right)+e_1e_2\mathbf{u}_S\Bigg)\mathbf{u}_S\\
=&\Big(\text{float}\left(2c_1+2c_2+4\right)\mathbf{u}+\text{float}\left(10c_1+10c_2+3d_1+3d_2+c_1{\cdot}c_2+97\right)\mathbf{u}^2\Big)\,\text{fl}\left(f_1{\cdot}f_2\right)\\
&+\frac{1}{(1+\mathbf{u})^4}\Bigg((1+\mathbf{u})^4\left(3-\mathbf{u}^2\right)f_1e_2+(1+\mathbf{u})^4\left(3-\mathbf{u}^2\right)f_2e_1\\
&+(1+\mathbf{u})^4\frac{1}{2}\left(6-6\mathbf{u}^2+\mathbf{u}^4\right)+(1+\mathbf{u})^4e_1e_2\mathbf{u}_S\Bigg)\mathbf{u}_S\\
\le&\Big(\text{float}\left(2c_1+2c_2+4\right)\mathbf{u}+\text{float}\left(10c_1+10c_2+3d_1+3d_2+c_1{\cdot}c_2+97\right)\mathbf{u}^2\Big)\,\text{fl}\left(f_1{\cdot}f_2\right)\\
&+\frac{1}{(1+\mathbf{u})^4}\Bigg((1+\mathbf{u})^4\left(3-\mathbf{u}^2\right)\left(\text{fl}\left(f_1{\cdot}e_2\right)+\frac{\mathbf{u}_S}{2}\right)+(1+\mathbf{u})^4\left(3-\mathbf{u}^2\right)\left(\text{fl}\left(f_2{\cdot}e_1\right)+\frac{\mathbf{u}_S}{2}\right)\\
&+(1+\mathbf{u})^4\frac{1}{2}\left(6-6\mathbf{u}^2+\mathbf{u}^4\right)+(1+\mathbf{u})^4(1+\mathbf{u})\text{fl}\left(e_1{\cdot}e_2\right)\mathbf{u}_S\Bigg)\mathbf{u}_S\\
=&\Big(\text{float}\left(2c_1+2c_2+4\right)\mathbf{u}+\text{float}\left(10c_1+10c_2+3d_1+3d_2+c_1{\cdot}c_2+97\right)\mathbf{u}^2\Big)\,\text{fl}\left(f_1{\cdot}f_2\right)\\
&+\frac{1}{(1+\mathbf{u})^4}\Bigg((1+\mathbf{u})^4\left(3-\mathbf{u}^2\right)\text{fl}\left(f_1{\cdot}e_2\right)+(1+\mathbf{u})^4\left(3-\mathbf{u}^2\right)\text{fl}\left(f_2{\cdot}e_1\right)\\
&+\frac{(1+\mathbf{u})^4}{2}\left(6-6\mathbf{u}^2+\mathbf{u}^4+\left(6-2\mathbf{u}^2\right)\mathbf{u}_S\right)+(1+\mathbf{u})^5\text{fl}\left(e_1{\cdot}e_2\right)\mathbf{u}_S\Bigg)\mathbf{u}_S\\
\le&\Big(\text{float}\left(2c_1+2c_2+4\right)\mathbf{u}+\text{float}\left(10c_1+10c_2+3d_1+3d_2+c_1{\cdot}c_2+97\right)\mathbf{u}^2\Big)\,\text{fl}\left(f_1{\cdot}f_2\right)\\
&+\frac{1}{(1+\mathbf{u})^4}\Bigg((1+\mathbf{u})^4\left(3-\mathbf{u}^2\right)\text{fl}\left(f_1{\cdot}e_2\right)+(1+\mathbf{u})^4\left(3-\mathbf{u}^2\right)\text{fl}\left(f_2{\cdot}e_1\right)
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{(1+\mathbf{u})^4}{2}\left(6-6\mathbf{u}^2+\mathbf{u}^4+\left(6-2\mathbf{u}^2\right)\mathbf{u}_S\right)+(1+\mathbf{u})^5\left(\mathrm{fl}\big((e_1\cdot e_2)\,\mathbf{u}_S\big)+\frac{\mathbf{u}_S}{2}\right)\Bigg)\mathbf{u}_S\\
=&\Big(\mathrm{float}\,(2c_1+2c_2+4)\,\mathbf{u}+\mathrm{float}\,(10c_1+10c_2+3d_1+3d_2+c_1\cdot c_2+97)\,\mathbf{u}^2\Big)\mathrm{fl}\,(f_1\cdot f_2)\\
&+\frac{1}{(1+\mathbf{u})^4}\Bigg((1+\mathbf{u})^4\left(3-\mathbf{u}^2\right)\mathrm{fl}\,(f_1\cdot e_2)+(1+\mathbf{u})^4\left(3-\mathbf{u}^2\right)\mathrm{fl}\,(f_2\cdot e_1)\\
&+\frac{(1+\mathbf{u})^4}{2}\left(6-6\mathbf{u}^2+\mathbf{u}^4+\left(7+\mathbf{u}-2\mathbf{u}^2\right)\mathbf{u}_S\right)+(1+\mathbf{u})^5\mathrm{fl}\big((e_1\cdot e_2)\,\mathbf{u}_S\big)\Bigg)\mathbf{u}_S\\
\le&\Big(\mathrm{float}\,(2c_1+2c_2+4)\,\mathbf{u}+\mathrm{float}\,(10c_1+10c_2+3d_1+3d_2+c_1\cdot c_2+97)\,\mathbf{u}^2\Big)\mathrm{fl}\,(f_1\cdot f_2)\\
&+\frac{1}{(1+\mathbf{u})^4}\Bigg(\mathrm{fl}\,(3+16\mathbf{u})\,\mathrm{fl}\,(f_1\cdot e_2)+\mathrm{fl}\,(3+16\mathbf{u})\,\mathrm{fl}\,(f_2\cdot e_1)\\
&+\frac{(1+\mathbf{u})^4}{2}\left(6-6\mathbf{u}^2+\mathbf{u}^4+\left(7+\mathbf{u}-2\mathbf{u}^2\right)\mathbf{u}_S\right)+\mathrm{fl}\,(1+6\mathbf{u})\,\mathrm{fl}\big((e_1\cdot e_2)\,\mathbf{u}_S\big)\Bigg)\mathbf{u}_S\\
\le&\Big(\mathrm{float}\,(2c_1+2c_2+4)\,\mathbf{u}+\mathrm{float}\,(10c_1+10c_2+3d_1+3d_2+c_1\cdot c_2+97)\,\mathbf{u}^2\Big)\mathrm{fl}\,(f_1\cdot f_2)\\
&+\frac{1}{(1+\mathbf{u})^4}\Bigg((1+\mathbf{u})\mathrm{fl}\,((3+16\mathbf{u})\cdot(f_1\cdot e_2))+\frac{\mathbf{u}_S}{2}+(1+\mathbf{u})\mathrm{fl}\,((3+16\mathbf{u})\cdot(f_2e_1))+\frac{\mathbf{u}_S}{2}\\
&+\frac{(1+\mathbf{u})^4}{2}\left(6-6\mathbf{u}^2+\mathbf{u}^4+\left(7+\mathbf{u}-2\mathbf{u}^2\right)\mathbf{u}_S\right)+(1+\mathbf{u})\mathrm{fl}\left((1+6\mathbf{u})\cdot\big((e_1\cdot e_2)\,\mathbf{u}_S\big)\right)+\frac{\mathbf{u}_S}{2}\Bigg)\mathbf{u}_S\\
=&\Big(\mathrm{float}\,(2c_1+2c_2+4)\,\mathbf{u}+\mathrm{float}\,(10c_1+10c_2+3d_1+3d_2+c_1\cdot c_2+97)\,\mathbf{u}^2\Big)\mathrm{fl}\,(f_1\cdot f_2)\\
&+\frac{1}{(1+\mathbf{u})^4}\Bigg((1+\mathbf{u})\mathrm{fl}\,((3+16\mathbf{u})\cdot(f_1\cdot e_2))+(1+\mathbf{u})\mathrm{fl}\,((3+16\mathbf{u})\cdot(f_2\cdot e_1))\\
&+\frac{(1+\mathbf{u})^4}{2}\left(6-6\mathbf{u}^2+\mathbf{u}^4+\left(7+\frac{3}{(1+\mathbf{u})^4}+\mathbf{u}-2\mathbf{u}^2\right)\mathbf{u}_S\right)\\
&+(1+\mathbf{u})\mathrm{fl}\left((1+6\mathbf{u})\cdot\big((e_1\cdot e_2)\,\mathbf{u}_S\big)\right)\Bigg)\mathbf{u}_S\\
<&\Big(\mathrm{float}\,(2c_1+2c_2+4)\,\mathbf{u}+\mathrm{float}\,(10c_1+10c_2+3d_1+3d_2+c_1\cdot c_2+97)\,\mathbf{u}^2\Big)\mathrm{fl}\,(f_1\cdot f_2)\\
&+\frac{1}{(1+\mathbf{u})^4}\Bigg((1+\mathbf{u})\mathrm{fl}\,((3+16\mathbf{u})\cdot(f_1\cdot e_2))+(1+\mathbf{u})\mathrm{fl}\,((3+16\mathbf{u})\cdot(f_2\cdot e_1))\\
&+(1+\mathbf{u})4+(1+\mathbf{u})\mathrm{fl}\left((1+6\mathbf{u})\cdot\big((e_1\cdot e_2)\,\mathbf{u}_S\big)\right)\Bigg)\mathbf{u}_S\\
=&\Big(\mathrm{float}\,(2c_1+2c_2+4)\,\mathbf{u}+\mathrm{float}\,(10c_1+10c_2+3d_1+3d_2+c_1\cdot c_2+97)\,\mathbf{u}^2\Big)\mathrm{fl}\,(f_1\cdot f_2)\\
&+\frac{1}{(1+\mathbf{u})^3}\Bigg(\mathrm{fl}\,((3+16\mathbf{u})\cdot(f_1\cdot e_2))+\mathrm{fl}\,((3+16\mathbf{u})\cdot(f_2\cdot e_1))\\
&+4+\mathrm{fl}\left((1+6\mathbf{u})\cdot\big((e_1\cdot e_2)\,\mathbf{u}_S\big)\right)\Bigg)\mathbf{u}_S\\
\le&\Big(\mathrm{float}\,(2c_1+2c_2+4)\,\mathbf{u}+\mathrm{float}\,(10c_1+10c_2+3d_1+3d_2+c_1\cdot c_2+97)\,\mathbf{u}^2\Big)\mathrm{fl}\,(f_1\cdot f_2)
\end{aligned}
$$

$$+\frac{(1+\mathbf{u})^3\mathrm{float}\Big((3+16\mathbf{u})\cdot(f_1\cdot e_2)+(3+16\mathbf{u})\cdot(f_2\cdot e_1)+4+(1+6\mathbf{u})\cdot\big((e_1\cdot e_2)\,\mathbf{u}_S\big)\Big)\mathbf{u}_S}{(1+\mathbf{u})^3}$$

$$\begin{aligned}
=&\Big(\mathrm{float}\,(2c_1+2c_2+4)\,\mathbf{u}+\mathrm{float}\,(10c_1+10c_2+3d_1+3d_2+c_1\cdot c_2+97)\,\mathbf{u}^2\Big)\mathrm{fl}\,(f_1\cdot f_2)\\
&+\mathrm{float}\Big((3+16\mathbf{u})\cdot(f_1\cdot e_2)+(3+16\mathbf{u})\cdot(f_2\cdot e_1)+4+(1+6\mathbf{u})\cdot\big((e_1\cdot e_2)\,\mathbf{u}_S\big)\Big)\mathbf{u}_S\\
\leq&\Big(\mathrm{float}\,(2c_1+2c_2+4)\,\mathbf{u}+\mathrm{float}\,(10c_1+10c_2+3d_1+3d_2+c_1\cdot c_2+97)\,\mathbf{u}^2\Big)\mathrm{fl}\,(f_1\cdot f_2)\\
&+\Big\lceil\mathrm{float}\Big((3+16\mathbf{u})\cdot(f_1\cdot e_2)+(3+16\mathbf{u})\cdot(f_2\cdot e_1)+4+(1+6\mathbf{u})\cdot\big((e_1\cdot e_2)\,\mathbf{u}_S\big)\Big)\Big\rceil\,\mathbf{u}_S
\end{aligned}$$

□

参考文献

[1] IEEE Computer Society : IEEE 754-2008 Standard for Floating-Point Arithmetic, New York, 2008.

[2] IEEE Computer Society : IEEE 754-1985 Standard for Binary Floating-Point Arithmetic, IEEE Computer Society, Washington, DC, 1985.

[3] Berg M., Cheong O., Kreveld M., Overmars M., 浅野哲夫 訳 : コンピュータ・ジオメトリ, 計算幾何学:アルゴリズムと応用 第 3 版, 近代科学社, 東京, 2010.

[4] Brönnimann, H., Burnikel, C., Pion, S. : Interval arithmetic yields efficient dynamic filters for computational geometry, Discrete Applied Mathematics, **109**, 25–47, (2001).

[5] Burnikel, C., Funke, S., Seel, M. : Exact Geometric Computation Using Cascading, International Journal of Computational Geometry & Applications, **11**, 245–266, (2000).

[6] Higham, N. J. : Accuracy and Stability of Numerical Algorithms, second edition, SIAM, Philadelphia, 2002.

[7] 樋口裕幸, 尾崎克久 : 浮動小数点演算による内積の丸め誤差解析, 日本応用数理学会論文誌, **26** 2, 182–212, (2016).

[8] Jeannerod, C.-P., Rump, S. M. : Improved error bounds for inner products in floating-point arithmetic, SIAM Journal on Matrix Analysis and Applications, **34**, 338–344, (2013).

[9] Jeannerod, C.-P, Rump, S. M, On relative errors of floating-point operations: optimal bounds and applications, Mathematics of Computation, **87**, 803–819, (2018).

[10] Kahan, W. M., Darcy, J. D. : How Java's Floating-Point Hurts Everyone Everywhere, ACM 1998 Workshop on Java for High-Performance Network Computing, 1998.
`http://www.cs.berkeley.edu/~wkahan/JAVAhurt.pdf`

[11] Kahan, W. M. : The regrettable failure of automated error analysis, a mini-course prepared for the conference at MIT on Computers and Mathematics, 1989.

[12] Kettner, L., Mehlhorn K., Pion, S., Schirra, S., Yap, C. : Classroom Examples of Robustness Problems in Geometric Computation, Computational Geometry, **40**, 61–78, (2007).

[13] Melquiond, G., Pion, S. : Formally Certified Floating-Point Filters For Homogeneous Geometric Predicates, RAIRO - Theoretical Informatics and Applications, **41**, 57–69, (2007).

[14] Muller, J.-M., Brisebarre, N., de Dinechin, F., Jeannerod, C.-P., Lefèvre, V.,

Melquiond, G., Revol, N., Stehlé, D., Torres, S. : Handbook of Floating-Point Arithmetic, Birkhäuser Boston, New York, 2010.

[15] Mohr, P. J., Newell, D. B., Taylor, B. N., Tiesinga, E. : Data and analysis for the CODATA 2017 special fundamental constants adjustment, Metrologia, **55** 1, (2018).

[16] 中尾充宏, 渡部善隆：実例で学ぶ精度保証付き数値計算:理論と実装, サイエンス社, 東京, 2011.

[17] Ogita, T., Rump, S. M., Oishi, S. : Verified Solutions of Linear Systems without Directed Rounding, Technical Report 2005-04, Advanced Research Institute for Science and Engineering, Waseda University, 2005.

[18] 大石進一：精度保証付き数値計算, コロナ社, 東京, 2000.

[19] 大石進一：応用解析セミナー 数値解析, 裳華房, 東京, 2004.

[20] O'Rourke, J. : Computational Geometry in C, Cambridge University Press, New York, 1998.

[21] Ozaki, K., Ogita, T., Rump, S. M., Oishi, S. : Adaptive and Efficient Algorithm for 2D Orientation Problem, Japan Journal of Industrial and Applied Mathematics, **26**, 215–231, (2009).

[22] Ozaki, K., Bünger, F., Ogita, T., Oishi, S., Rump, S. M. : Simple floating-point filters for the two-dimensional orientation problem, BIT Numerical Mathematics, **56** 2, 729–749, (2015).

[23] Pan, V. Y., Yu. Y. : Certified Numerical Computation of the Sign of a Matrix Determinant, Algorithmica, **130**, 708–724, (2001).

[24] Rump, S. M. : Algorithms for Verified Inclusions—Theory and Practice. In R. E. Moore, editor, Reliability in Computing: The Role of Interval Methods in Scientific Computing, Academic Press Professional, Inc. , San Diego, 109–126, (1998).

[25] Rump, S. M. : Verification methods: Rigorous results using floating-point arithmetic, Acta Numerica, **19**, 287–449, (2010).

[26] Rump, S. M. : Error estimation of floating-point summation and dot product, BIT Numerical Mathematics, **52** 1, 201–220, (2012).

[27] Rump, S. M. : Computable backward error bounds for basic algorithms in linear algebra, Nonlinear Theory and Its Applications, IEICE, **6**, 360–363, (2015).

[28] Rump, S. M., Ogita, T., Oishi, S. : Accurate floating-point summation part I: Faithful rounding, SIAM Journal of Scientific Computing, **31**, 189–224, (2008).

[29] Shewchuk, J. R. : Adaptive precision floating-point arithmetic and fast robust geometric predicates, Discrete & Computational Geometry, **18**, 305–363, (1997).

[30] Salesin, D., Stolfi, J, Guibas, L. : Epsilon Geometry: Building Robust Algorithms

from Imprecise Computations, SCG '89 Proceedings of the fifth annual symposium on Computational geometry, New York, 208–217, (1989).

[31] 杉原厚吉：計算幾何学, 朝倉書店, 東京, 2013.

[32] 譚学厚, 平田富夫： 計算幾何学入門: 幾何アルゴリズムとその応用, 森北出版, 東京, 2001.

[33] Winograd, S. : A New Algorithm for Inner Product, IEEE Transactions on Computers, **17**, 693–694, (1986).

研究業績書

原著論文

[R1] 太田悠暉, 尾崎克久： 点と有向直線の位置関係に対する浮動小数点フィルタの実数入力への拡張と凸包への応用, 日本応用数理学会論文誌, **24** 4, 373–395, (2014).

[R2] 太田悠暉, 尾崎克久： 計算値の大小関係を保証する浮動小数点フィルタ, 日本応用数理学会論文誌, **28** 1, 1–17, (2018).

査読付き国際会議論文

[R3] Ohta, Y., Ozaki, K.： Computable error bounds for floating-point filters, Proceeding of The 37th JSST Annual International Conference on Simulation Technology, 99–102, (2018).

[R4] Ohta, Y., Ozaki, K.： Extension of floating-point filters to absolute and relative errors for numerical computation, The International Conference on Mathematics: Pure, Applied and Computation (ICoMPAC) 2018, to appear in IOP Conference Proceeding.

国内発表

[R5] 太田悠暉, 尾崎克久： 区間入力に対する幾何判定問題の精度保証化に関する準備, 2012, 日本応用数理学会 2012 年度 年会, (2012/08/31).

[R6] 太田悠暉, 尾崎克久： 浮動小数点で近似されたデータに対する凸包の精度保証アルゴリズムについて, 2013, 日本応用数理学会 第 9 回 研究部会連合発表会, (2013/03/15).

[R7] 太田悠暉, 尾崎克久： 区間入力まで拡張した Orient2D の浮動小数点フィルタと凸包構成への適用, 2015, 日本応用数理学会 環瀬戸内研究部会 第 18 回 環瀬戸内ワークショップ, (2015/09/25).

[R8] 太田悠暉, 尾崎克久： 2 次元平面における 2 点間の距離の大小判定問題に関する精度保証法, 2016, 日本応用数理学会 2016 年度 年会, (2016/09/14).

[R9] 太田悠暉, 尾崎克久： 計算値の大小関係を保証する浮動小数点フィルタについて, 2017, 日本応用数理学会 第 13 回 研究部会連合発表会, (2017/03/07).

[R10] 太田悠暉, 尾崎克久： 計算値の大小関係を保証する浮動小数点フィルタ, 2017, 日本応用数理学会 環瀬戸内研究部会 第 26 回 環瀬戸内ワークショップ, (2017/07/22).

[R11] 太田悠暉, 尾崎克久：2 数の大小比較に対する浮動小数点フィルタの生成方法について, 2017, 第 1 回 精度保証付き数値計算の実問題への応用研究集会, (2017/12/09).

[R12] 太田悠暉, 尾崎克久：浮動小数点フィルタの応用と性能評価, 2018, 日本応用数理学会 第 14 回 研究部会連合発表会, (2018/03/15).

国際会議

[R13] Ohta, Y., Ozaki, K. : Convex Hull for Set of Real Numbers with Verified Numerical Computation, 2013, The 9th East Asia SIAM Conference, (2013/06/19).

[R14] Ohta, Y., Ozaki, K. : 2D Orientation Problem for Interval Data, 2013, The 32th JSST Annual Conference: International Conference on Simulation Technology, (2013/09/13).

[R15] Ohta, Y., Ozaki, K. : Iterative Convex hull Algorithm with Verified Numerical Computation, 2014, The 10th East Asia SIAM Conference, (2014/06/24).

[R16] Ohta, Y., Ozaki, K. : Verified Convex Hull for Inexact Data, 2015, 8th Small Workshop on Interval Methods, (2015/06/10).

[R17] Ohta, Y., Ozaki, K. : Verification of Distances in Two-Dimensional Space, 2016, 5th European Seminar on Computing, (2016/06/09).

[R18] Ohta, Y., Ozaki, K. : Iterative algorithm for convex hull based on floating-point filters, 2016, Czech-Japanese-Polish Seminar in Applied Mathematics 2016, (2016/09/07).

[R19] Ohta, Y., Ozaki, K. : Iterative algorithms based on verification methods for computational geometry, 2017, The International Workshop on Numerical Verification and its Applications 2017, (2017/03/17).

[R20] Ohta, Y., Ozaki, K. : Applications of Floating-Point Filters, 2018, IX Pan-American Workshop Applied Mathematics & Computational Science, (2018/06/14).

[R21] Ohta, Y., Ozaki, K. : Computable error bounds for floating-point filters, 2018, The 37th JSST Annual Conference: International Conference on Simulation Technology, (2018/09/18).

[R22] Ohta, Y., Ozaki, K. : Extension of floating-point filters to absolute and relative errors for numerical computation, The International Conference on Mathematics: Pure, Applied and Computation (ICoMPAC) 2018, (2018/10/20).